Statistical Models for Ordination and Biodiversity Analysis in Community Ecology

Yingjie Zhang

Dissertation submitted in fulfilment of the requirements for the degree of

Doctor in Statistical Data Analysis

Academic year 2016 - 2017

Promoters:

Prof. dr. ir. Olivier Thas

Department of Mathematical Modelling, Statistics, and Bioinformatics, Faculty of Bioscience Engineering, Ghent University.

Prof. dr. Jan De Neve

Department of Data Analysis, Faculty of Psychology and Educational Sciences, Ghent University.

Dean:

Prof. dr. Herwig Dejonghe

Rector:

Prof. dr. Anne De Paepe

Please refer to this work as follows:

Yingjie Zhang (2016). Statistical models for ordination and biodiversity analysis in community ecology, PhD thesis, Ghent University, Belgium

ISBN: 978-90-5989-939-1

To my father

Acknowledgements

Journey to the West is an epic tale of the historical journey of a monk, who traveled all the way to India in order to bring the holy Mahayana Buddhist scriptures back to China. In the monk's journey, the evil creatures sought to kill and eat him because his flesh is said to impart immortality. I often compare my journey of completing the PhD to the monk's journey, it is as adventurous but surly less dramatic. The holy Mahayana Buddhist scripture I attained from this journey is this dissertation, I want to thank many people for accompanying me in this journey.

I would like to first say a very big thank you to my promotor Prof. Olivier Thas for all the support he has given me. In 2010, he encouraged me to start a PhD which I have never thought I could be. Thank you Olivier for being so patient, encouraging, inspirational and helpful. My deep appreciation also goes to my promoter Prof. Jan De Neve for his great contribution to Chapter 4 of my thesis and the support he gave. I would also like to thank the examination committees for the suggestions and remarks in improving the dissertation.

I gratefully acknowledge the funding received towards my PhD from the project of Linking Sino-European Universities through Mobility (LiSUM) and the IAP research network grant P7/06 of the Belgian Science Policy and by the Multidisciplinary Research Partnership Bioinformatics: from nucleotides to networks of Ghent University.

I would also like to say thank you to the colleagues from the Department of Mathematical Modelling, Statistics and Bioinformatics for the lovely work environment. A special thank you goes to Guillermo Vidal, Aisling Daly, Marc Sader for helping with organizing the reception and for being great friends. I am grateful with the wonderful officemates

I had in A1.049, thanks for your kindness and contribution to the relaxing moments.

Finally my great appreciation to the family who have been by my side throughout this PhD. A special thank you goes to my husband Guangchao, without him it was not possible to complete the PhD.

Samenvatting

In gemeenschapsecologie worden ondermeer abiotische relaties tussen gemeenschappen en omgevingsfactoren bestudeerd. Hiertoe worden op verschillende locaties abundanties van planten- of dierensoorten geobserveerd, alsook de locatie-specifieke omgevingsfactoren. Ook microbiële gemeenschappen kunnen bestudeerd worden: de abundanties van micro-organismen kunnen bekomen worden via de sekwenering van het 16S rRNA gen.

Het bestuderen van de relaties tussen abundanties en omgevingsfactoren gebeurt dikwijls via ordinatiemethoden (bv. canonische correspondentie-analyse, CCA). De klassieke ordinatiemethoden kunnen echter misleidende resultaten genereren voor datasets waarin veel nul-abundanties voorkomen. Dergelijke nul-abundanties komen veelvuldig voor in abundantiestudies, en in het bijzonder in studies van microbiële gemeenschappen die getypeerd worden door de hele vele species die gedetecteerd kunnen worden. Om dit probleem te verhelpen, hebben we in hoofdstuk 2 een nieuwe ordinatiemethode ontwikkeld. De methode is gebaseerd op een model-gebaseerde ordinatie methode, waarin we de *zero-inflated Poisson* (ZIP), de *zero-inlfated negative binomial* (ZINB) of *hurdle* modellen integreren om de hoge frequentie aan nullen op te vangen. Simulatiestudies bevestigen dat onze methoden een correcter resultaat geven dan de reeds bestaande methoden. Een ander zwak punt van de klassieke CCA is dat deze impliciet veronderstelt dat een species-responsfunctie, die uitdrukt hoe de verwachte abundantie varieert met een univariate omgevingsscore die door de methode bepaald wordt als een lineaire combinatie van de gemeten omgevingsfactoren,

unimodaal is en een maximum vertoont (klokvorm). Voor ieder species wordt dit maximum aangeduid in de ordinatiefiguur. Responsfuncties moeten klokvormig zijn volgens de niche-theorie uit de ecologie. Model-gebaseerde ordinatiemethoden bieden echter geen garantie dat de geschatte responsfuncties een maximum vertonen (het kan ook een minimum zijn, i.e. de responsfunctie heeft een U-vorm), waardoor de punten in de ordinatiefiguren niet noodzakelijk geïnterpreteerd kunnen worden als maxima. In hoofdstuk 3 hebben we een nieuwe model-gebaseerde ordinatiemethode ontwikkeld die U-vormen penaliseert. De methode geeft een oplossing met meer klokvormige responsfuncties zodat de ordinatiefiguur een meer informatieve interpretatie toelaat.

Gemeenschappen worden dikwijls samengevat door een biodiversiteitsindex, zodat verbanden tussen omgevingsfactoren en de biodiversiteit bestudeerd kunnen worden. Dergelijke analyses worden in twee fases uitgevoerd: (1) een biodiversiteitsindex wordt berekend voor iedere locatie; (2) deze biodiversiteitsindices worden als uitkomst beschouwd in een regressiemodel waarin de omgevingsfactoren als regressoren opgenomen worden. Deze methode houdt echter geen rekening met de varantieheterogeniteit van de biodiversiteitsschattingen. In deze thesis beschouwen we de Gini index als biodiversiteitsindex. In hoofdstuk 4 hebben we een semiparametrisch model ontwikkeld voor het analyseren van het effect van een regressor (bv. omgevingsfactor) op de gemiddelde Gini index. De theoretische ontwikkelingen zijn gebaseerd op beperkende veronderstellingen en verder onderzoek is nodig om de methode ruimer toepasbaar te maken.

Doorheen de hele thesis worden vier datasets gebruikt ter motivatie en voor de illustratie van de methoden. Twee datasets komen van traditionele abundantiestudies van dieren (spinnen en mijten) en één dataset komt van een microbiële metagenomics

studie: microbiële gemeenschappen in Antarctische meren. In hoofdstuk 5 worden de methoden uit hoofdstukken 2 en 4 toegepast op een longitudinale humane microbioomstudie van jonge kinderen met aanleg voor type I diabetes. De omgevingsfactoren bestaan hier uit informatie over het dieet van de kinderen. Dit hoofdstuk illustreert dat de methoden ontwikkeld in dit doctoraatsonderzoek ook een meerwaarde kunnen betekenen voor microbioomstudies.

Contents

List of abbreviations

α-diversity	alpha diversity
β-diversity	beta diversity
γ-diversity	gamma diversity
aSSE	average SSE
BECOA	Bell-shape Enriched COA
CA	Correspondence analysis
CCA	Canonical correspondence analysis
COA	Constrained Ordination Analysis
CGO	Constrained Gaussian Ordination
FCOA	Flexible Constrained Ordination Analysis
GEE	generalised estimating equation
GLMs	generalised linear model
HZTP	hurdle-zero-truncated-Poisson
LDA	Fisherś linear discriminant analysis
LR	log-likelihood ratio
OTU	Operational Taxonomic Unit
PCR	Polymerase Chain Reactions
rRNA	ribosomal RNA
SSE	the total sum of squared errors
T1D	type 1 diabetes
ZAD	zero-altered distributions
ZIP	zero- inflated Poisson
ZINB	zero-inflated negative binomial
ZTP	zero-truncated Poisson
ZTNB	zero-truncated negative binomial

Chapter 1

Introduction

1.1 Background

The core of ecology is to discover modes of interaction among a community of living organisms and their environment. A living organism can interact with other organisms living in the community and this type of interaction is called biotic interaction [**?**]. On the other hand, an organism can also interact with chemical and physical parts of the environment, and this is called abiotic interaction. The abiotic interaction with local environmental conditions in this context is actually twofold. One type of interaction refers to the role that environmental conditions play in the spatial and temporal distribution of individual species. Knowing this interaction is helpful to find out which environmental conditions elevate growth or existence of one or more species, or, on the contrary, which environmental factors may lead to extinction. The other type of abiotic interaction is concerned with the community level, particularly the relationship between the habitat conditions and the community composition. For instance, biodiversity may be affected

by local environmental conditions.

The scope of species-community studies spans from the planetary scale to the micro-level, depending on the size of the study subjects. Microecology or microbial ecology is the subfield that copes with microorganisms such as bacteria and fungi, whereas macroecology focuses on plants and animals. Before the availability of good microscope technologies, macroecology encompassed the study of environmental changes and how they affect the wildlife population. For example, how does a rise in temperature changes the distribution of birds in Europe? To understand the complex mechanism of the ecosystem is a great challenge. In this thesis, the focus is on some aspects of the abiotic interaction.

In the early years of microbial ecology, scientists could only identify microbial species through culture-dependent techniques, i.e., microbial communities could only be identified and quantified if all of their constituting microorganisms could grow in laboratory conditions without affecting their relative abundances. Many microorganisms, however, do not grow under these laboratory conditions, limiting the study of complex communities. In the early years of the 21st century, the innovation of high-throughput sequencing techniques opened the door for a more detailed analysis of microbial communities. The sequencing techniques no longer require the cultivation of microbial community samples in laboratories. A small sample is sufficient; the sequencing techniques start from DNA extracted from the microbial cells. The identification of tens of thousands of microbes in a sample became much cheaper and more efficient. This has led to a new research area called *metagenomics*, where high-throughput sequencing is used for species identification and quantification of relative abundances. This technique will be given a closer look later in this chapter. However, the size of the data sets resulting

from metagenomics studies can be challenging for statistical methods that were originally designed for small to moderately sized data sets. In general, metagenomics starts from a small sample taken from a particular location. The sample may, for example, be an environmental sample taken from soil or a water body (environmental genomics or ecogenomics). However, samples may also be taken from human body sites, which is the subject of microbiome studies.

In both macroecology and microecology, data on the community structure in a given habitat often consist of abundances or relative abundances of a collection of species. For each habitat, or sampling location, the researchers also collect data on the environmental conditions. In macroecology the sampling locations refer to geographical locations, whereas in microbial ecology the sampling locations may also refer to humans or even specific organs of humans. With the data on the abundances and on the environmental conditions, researchers may be interested in finding relationships between the environment and the abundances: which species favour what environment? Ordination methods form a popular class of data analysis methods to answer these research questions. Or the researchers may aggregate the abundances into a biodiversity measure as a summary statistic of community structure, and study how the biodiversity is affected by the environment.

In this thesis we propose new ordination methods for studying the relationship between environment and species abundances. Our ordination methods are designed to improve on existing methods with respect to some issues that arise in many abundance studies (e.g. zero abundances). We also propose a semiparametric regression model for analysing biodiversity as a function of covariates (e.g. environmental conditions). The rest of this chapter is organised as follows. An introduction to classical ordination

analysis is given in Section 1.2. This section also includes a description of the typical data structure of abundance studies, as well as an illustration of a classical ordination analysis and its visualisation. Section 1.3 gives an introduction to species diversity indices. Several data sets are used throughout the thesis; these are described in Section 1.4. Since two metagenomics data sets are included, this section also contains a brief introduction to the technology behind the data generation. Finally, in Sections 1.5 and 1.6 we discuss some of the outstanding challenges in community ecology, and we formulate the research objectives of this thesis.

1.2 Ordination analysis

1.2.1 Data structure

The data sets collected for studying the association between either a macroscopic or microscopic community and their habitats often come from field surveys or from biological samples. In field surveys, samples are first taken from sampling sites according to a particular design, while the environmental measurements of the habitat are also recorded. Examples of the latter are: pH, temperature, humidity, or concentrations of chemical compounds in the environment. Human biological samples can also be obtained from different body sites: e.g., from scraping the skin, a faeces sample, or from a biopsy. In human microbiome studies, baseline characteristics of the subject (e.g., age, gender), data on health status, and local environmental conditions (e.g., pH of the skin), may be collected for explaining the variability in the microbial community composition.

For each individual sample species or their taxonomic groups are identified and quantified, either through classical methods or modern high-throughput sequencing technologies. The resulting information is often orginized into two data objects.

The first data object consists of information on the species abundances. We denoted this abundance data matrix as $\boldsymbol{Y} = \{y_{ik}\}$ where y_{ik} is the abundance of species k at sampling location i, and where $i = 1, \ldots, n$ refers to the n sampling locations or sites and $k = 1, \ldots, K$ to the K species observed in the experiment. The other data object stores the information on the environmental measurements. It is denoted as the environmental data matrix $\boldsymbol{X} = \{x_{ij}\}$ where the n rows $i = 1, \ldots, n$ correspond to the same sampling sites as $\boldsymbol{Y}$, and $j = 1, \ldots, p$ refers to the p environmental variables. Note that we will use the term "environmental variable" generically throughout this thesis: it may also be used to refer to e.g. the baseline characteristics of humans in microbiome studies. Figure 1.1 shows a schematic overview of the data structure.

When the abundance is to be considered as a random variable, we will use the notation Y_{ik}.

$\boldsymbol{Y}=$

	species		
	1	...	K
site 1	y_{11}	...	y_{1K}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
site n	y_{n1}	...	y_{nK}

$\boldsymbol{X}=$

	measurement		
	1	...	p
site 1	x_{11}	...	x_{1p}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
site n	x_{n1}	...	x_{np}

Figure 1.1: Overview of the data structures of the abundance matrix $\boldsymbol{Y}$ and the environmental data matrix $\boldsymbol{X}$.

Tables 1.1 and 1.2 are example abundance and environmental matrices from the Dutch

dune hunting spider data set [**?**]. A detailed introduction of the data set is given later in Section 1.4.

Site	Alopacce	Alopcune	Alopfabr	Arctlute	Arctperi	Auloalbi	Pardlugu	Pardmont	Pardnigr	Pardpull	Trocterr	Zoraspin
1	25	10	0	0	0	4	0	60	12	45	57	4
2	0	2	0	0	0	30	1	1	15	37	65	9
3	15	20	2	2	0	9	1	29	18	45	66	1
4	2	6	0	1	0	24	1	7	29	94	86	25
5	1	20	0	2	0	9	1	2	135	76	91	17
6	0	6	0	6	0	6	0	11	27	24	63	34
7	2	7	0	12	0	16	1	30	89	105	118	16
8	0	11	0	0	0	7	55	2	2	1	30	3
9	1	1	0	0	0	0	0	26	1	1	2	0
10	3	0	1	0	0	0	0	22	0	0	1	0
11	15	1	2	0	0	1	0	95	0	1	4	0
12	16	13	0	0	0	0	0	96	1	8	13	0
13	3	43	1	2	0	18	1	24	53	72	97	22
14	0	2	0	1	0	4	3	14	15	72	94	32
15	0	0	0	0	0	0	6	0	0	0	25	3
16	0	3	0	0	0	0	6	0	2	0	28	4
17	0	0	0	0	0	0	2	0	0	0	23	2
18	0	1	0	0	0	0	5	0	0	0	25	0
19	0	1	0	0	0	0	12	0	1	0	22	3
20	0	2	0	0	0	0	13	0	0	0	22	2
21	0	1	0	0	0	0	16	1	0	1	18	2
22	7	0	16	0	4	0	0	2	0	0	1	0
23	17	0	15	0	7	0	2	6	0	0	1	0
24	11	0	20	0	5	0	0	3	0	0	0	0
25	9	1	9	0	0	2	1	11	6	0	16	6
26	3	0	6	0	18	0	0	0	0	0	1	0
27	29	0	11	0	4	0	0	1	0	0	0	0
28	15	0	14	0	1	0	0	6	0	0	2	0

Table 1.1: The abundance matrix of the hunting spider data set. The rows represent sampling locations and the columns represent twelve spider species. The abbrevation of the species names are: Alopacce (*Alopecosa accentuata*), Alopfabr (*Alopecosa fabrilis*), Arctlute (*Arctosa lutetiana*), Arctperi (*Arctosa perita*), Auloalbi (*Aulonia albimana*), Pardlugu (*Pardosa lugubris*), Pardmont (*Pardosa monticola*), Pardnigr (*Pardosa nigriceps*), Pardpull (*Pardosa pullata*), Trocterr (*Trochosa terricola*) and Zoraspin (*Zora spinimana*).

Site	WaterCon	BareSand	FallTwig	CoveMoss	CoveHerb	ReflLux
1	2.33	0.00	0.00	3.04	4.45	3.91
2	3.05	0.00	1.79	1.10	4.56	1.61
3	2.56	0.00	0.00	2.40	4.61	3.69
4	2.67	0.00	0.00	2.40	4.62	3.00
5	3.02	0.00	0.00	0.00	4.62	2.30
6	3.38	2.40	3.43	2.40	3.43	0.69
7	3.18	0.00	0.00	0.69	4.62	2.30
8	2.62	0.00	4.26	1.10	3.43	0.69
9	2.48	0.00	0.00	4.33	3.26	3.40
10	2.20	3.93	0.00	3.43	3.04	3.69
11	2.22	0.00	0.00	4.11	3.71	3.69
12	2.29	0.00	0.00	3.83	4.03	3.69
13	3.52	1.79	1.79	0.69	4.51	3.40
14	3.09	0.00	0.00	1.79	4.56	1.10
15	3.27	0.00	4.39	0.69	3.04	0.69
16	3.03	0.00	4.61	0.69	0.69	0.00
17	3.33	0.00	4.45	0.69	3.04	1.10
18	3.12	0.00	4.39	0.00	3.04	1.10
19	2.92	0.00	4.51	1.61	1.61	0.00
20	3.11	0.00	4.60	0.69	0.69	0.00
21	2.98	0.00	4.56	0.69	1.79	0.00
22	1.25	3.26	0.00	4.33	0.69	3.91
23	1.19	3.04	0.00	4.03	3.26	4.09
24	1.65	3.26	0.00	4.03	3.04	4.01
25	1.82	3.58	0.00	1.10	4.11	2.30
26	0.99	4.51	0.00	1.79	1.79	4.38
27	0.96	2.40	0.00	3.83	3.43	3.69
28	0.96	3.43	0.00	3.71	3.43	3.69

Table 1.2: The environmental matrix of the hunting spider data set. The rows represent sampling locations and the columns represent six environmental variables. The environmental variables are *BareSand* (percent cover of bare sand), *CoveMoss* (percent cover of the moss layer), *CoveHerb* (percentage coverage of the herb layer), *FallTwig* (percentage coverage of fallen leaves and twigs), *ReflLux* (reflection of the soil surface with cloudless sky) and *WaterCon* (percentage of soil dry mass).

When the species are identified and their abundances quantified with sequencing technologies, the abundance matrix $\boldsymbol{Y}$ is often sparse, i.e., many of the matrix entries are zero. This is because many different species can be identified but only a few species are present in all samples. Moreover, rare species tend to be undetected (see Section 1.4.1 for more details).

1.2.2 Species response function

Each species exists, grows and reproduces in its preferred habitat which is characterised by many environmental factors. The habitat where a species occurs is often named an *ecological niche*. *The fundamental niche theory* states that there is one particular environmental condition in which the species is most successful in terms of abundance. Moving further away from this optimal environment causes a decrease in the expected abundance. In other words, within its niche, each species tends to be most abundant in its most preferred environment [**???**].

The relationship between the environment and the expected abundance is quantified through the *species response function*. We first introduce the *environmental score* for sampling location i, which is denoted by z_i. It is a univariate variable that represents an aspect of the environmental condition at location i; it is sometimes also referred to as the *ordination*. More details will be given later. A unimodal response curve for a particular species k is often represented as follows:

$$f_k(z_i) = \mathrm{E}\left(Y_{ik} \mid z_i\right) = g\left(a_k - \frac{(z_i - \mu_k)^2}{2t_k^2}\right), \tag{1.1}$$

where $g(\cdot)$ is a monotonic link function (e.g., the exponential function), a_k is the maximum expected abundance of species k, μ_k is the value of the environmental score at which the maximum expected abundance is reached (*optimum*), and t_k is the tolerance. A larger tolerance means the species is less sensitive to deviations from the optimal environmental condition. When $g(\cdot)$ is the exponential function, the species response function is also known as the Gaussian response function, this is illustrated in Figure 1.2. Note that the response function is *bell-shaped*, i.e., it is unimodal and shows a

maximum.

The environmental score z is often expressed as a linear combination of the environmental measurements x_{ij}. In particular, $z_i = \boldsymbol{\alpha}^T \boldsymbol{x}_i$, with $\boldsymbol{x}_i^T = (x_{i1}, \ldots, x_{ip})$ and $\boldsymbol{\alpha}$ is the *environmental gradient*. The elements of $\boldsymbol{\alpha}$ are referred to as the *loadings* or the *coefficients* that make up the environmental gradient. Meaningful exceptions to the unimodal species response function exist. For example, **???** discusses the use of bimodal response curves. Sometimes the species response function does not show a maximum. For example, when the expected abundance is studied as a function of a toxic compound, a monotonically decreasing function is expected. In this thesis the focus is on the unimodal response curve which is the backbone of many analytical methods and it is often ecologically meaningful.

We can write Equation (1.1) using a Generalised Linear Model (GLM) with a second order polynomial regression model,

$$g^{-1}(f_k(z_i)) = \beta_{0k} + \beta_{1k} z_i + \beta_{2k} z_i^2. \tag{1.2}$$

In contrast to the nonlinear model (1.1), the polynomial regression does only result in a bell-shape curve for negative β_{2k}.

The equivalence between the sets of parameters of model (1.1) and (1.2) can be easily derived,

$$\beta_{0k} = a_k - \frac{\mu_k^2}{2t_k^2}, \quad \beta_{1k} = \frac{\mu_k}{t_k^2}, \quad \beta_{2k} = -\frac{1}{2t_k^2}.$$

Parameter estimation in models (1.1) and (1.2) is part of our research and will be discussed in later chapters.

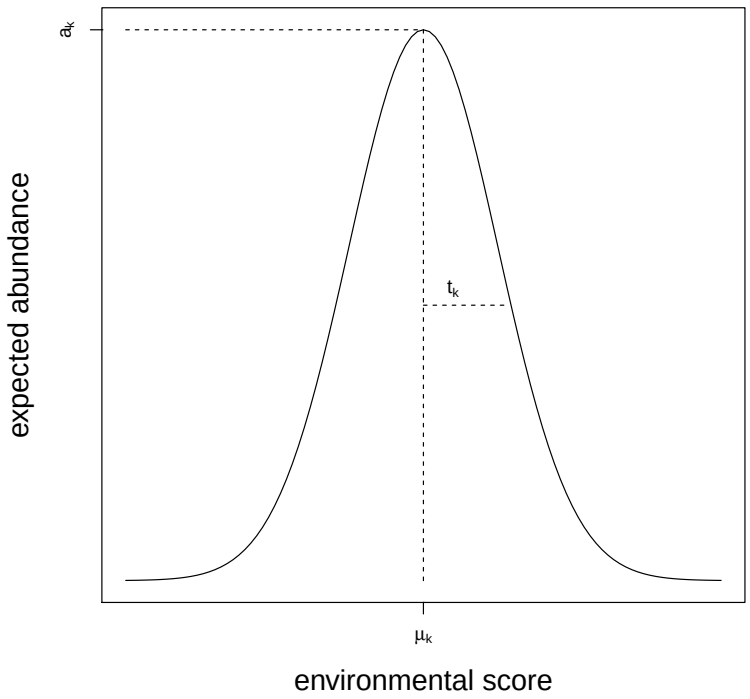


Figure 1.2: Gaussian species response function. a_k is the maximum abundance of species k, μ_k is the environmental score at which the abundance of species k reaches its maximum and t_k is the tolerance of species k.

Figure 1.3 displays the abundance data of species *Trocterr* from the hunting spider data. The left panel shows the data in a three-dimensional scatter plot, whereas, in the right panel, the fitted Gaussian response curve is plotted along an artificial environmental gradient.

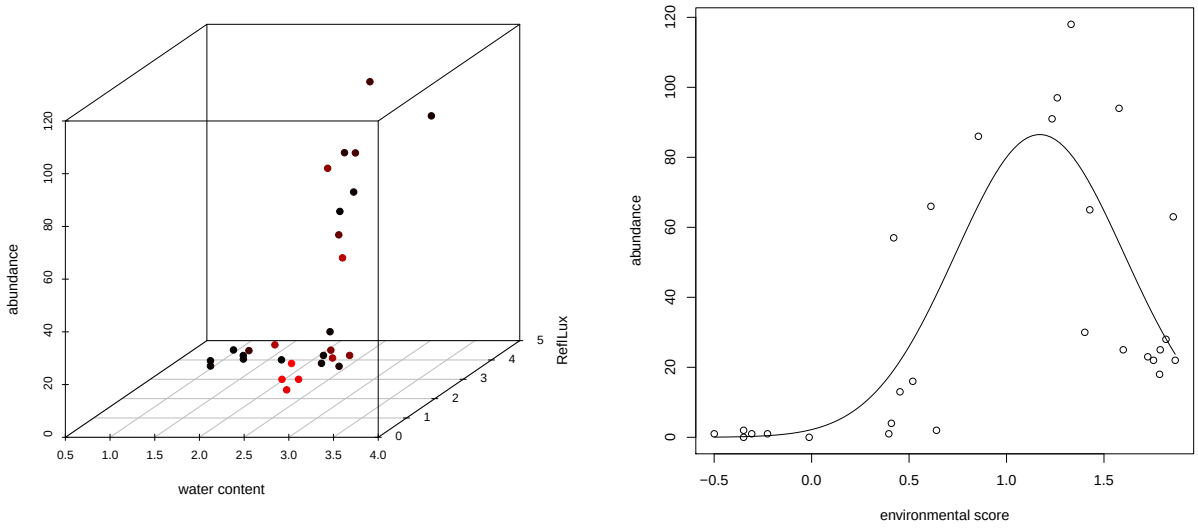


Figure 1.3: Abundance data of species *Trocterr* plotted against the environmental variables: *water* (water content) and *ReflLux* (Reflection of the soil surface with cloudless sky) (left) and the fitted species response curve of species *Trocterr* along an artificial environmental gradient (0.6× *water* - 0.25× *ReflLux*) (right). The dots correspond to the observed abundances.

1.2.3 A review of ordination analysis

Ordination analysis aims at studying the species-environment relationship by representing the original high-dimensional data in an informative low-dimensional graph. This dimension reduction comes at the cost of loss of information, but enables practitioners to visually explore the community structure in a lower dimensional space. The emergence of the term 'ordination' in ecology literature can be traced back to **?**, who introduced Principle Component Analysis for extracting ecologically meaningful informa-

tion from vegetation studies. Ever since a variety of papers on developing multivariate methods for ecological community data has been published; see e.g., **????**.

Ordination methods were merely considered as exploratory techniques due to lack of theoretical foundations until the rise of Correspondence Analysis (CA). For a recent account on CA we refer to **?**. **?** introduced CA under the name of *reciprocal averaging* to ecology. CA reveals a community structure by first arranging species and samples along latent variables (environmental gradients). **?** further proposed to relate the latent variables to some environmental measurements by a linear regression model so as to uncover the important environmental variables that affect the community structure. After dimension reduction, the new dimensions are also referred to as the *axes*; this name comes from the axes of the plot used to visualise the dimension reduced data. However, this two-step analysis has obvious limitations: an important environmental variable may not relate to the axes that forms first dimension but it can be still important to explain the relation between community composition and the environment [**?**]. Canonical Correspondence Analysis (CCA, **?**) brought a solution: it uses the environmental information directly in the ordination analysis (as a one-step method). This method has become a very widely used statistical tools in community ecology.

Correspondence analysis and canonical correspondence analysis

First we briefly outline the Correspondence Analysis (CA) approach. In CA each row of the abundance matrix $\boldsymbol{Y} = \{y_{ik}\}$ gets a site score assigned, denoted by z_i, and each column gets a species score assigned, denoted by μ_k. It can be shown that the latter corresponds to the optimum of a species response function. These scores are chosen

so that the ratio of a weighted sum of squares of the sample scores over a weighted sum of squares of the sample scores within species is maximised, i.e., the criterion to be maximised is

$$D^2 = \frac{\sum_i y_{i+}(z_i - r)^2}{\sum_k \sum_i y_{ik}(z_i - \mu_k)^2},$$

where $r = \sum_i y_{i+} z_i / y_{++}$ with $y_{i+} = \sum_{k=1}^{K} y_{ik}$ and $y_{++} = \sum_{i=1}^{n} y_{i+}$.

The estimates of μ_k and z_i can be obtained in an iterative manner. Maximisation of D^2 results in species scores close to the site scores where they are most abundant. In the next step the site scores are related to environmental variables.

Apart from the obvious limitation of two-steps methods, these procedures are also computationally demanding. **?** proposed Canonical Correspondence Analysis (CCA), which circumvents both limitations of CA under the assumption of unimodal species response curves. In particular, the following assumptions have to hold:

1. Equal maximum expected abundances, i.e., $a_1 = \ldots = a_k$.
2. Unit tolerances, i.e., $t_1 = \ldots = t_k = 1$.
3. The optima μ_k distribute homogeneously (i.e., site scores are equispaced or drawn randomly from a uniform distribution.) over an interval that is large in comparison to the species tolerance t_k.
4. Site scores z_i distribute uniformly over an interval that contains the all μ_k's.

Under conditions 1 - 4 ter Braak (**?**) showed that the two-step CA can be performed in one step. Part of his solution lies in relating the site scores z_i to the p-dimensional environmental variable $\boldsymbol{x}_i$ through the *environmental gradient* $\boldsymbol{\alpha}$, i.e., $z_i = \boldsymbol{\alpha}^t \boldsymbol{x}_i$, and replacing the direct estimation of the z_i by estimation of α.

The ordination methods described until now give one set of solutions (site and species scores, and the environmental gradient). The methods can be easily extended so that a second set of solutions is obtained. This is accomplished by requiring that the two sets of scores are orthogonal to one another. The first and the second set of scores are referred to as the first and the second dimension of the ordination analysis, and, similarly, in CCA the corresponding environmental gradients are referred to as the first and the second environmental gradients, or gradients of the first and second dimension. Plotting one dimension against the other, and using different symbols for site and species scores, gives an ordination plot. Several versions of such plots have been proposed. In CCA also the environmental gradients can be added to the graph. The graphs differ, e.g., in scaling of the two axes or of the two types of scores. Examples will be given in Section 1.2.4.

ter Braak's algorithm finds the first two environmental gradients $\boldsymbol{\alpha}$ as the first two eigenvectors of the matrix

$$(\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{C}^{-1}\boldsymbol{Y}^T\boldsymbol{X},$$

where $\boldsymbol{R} = \text{diag}\{y_{i+}\}$ and $\boldsymbol{C} = \text{diag}\{y_{+k}\}$, and $y_{+k} = \sum_{i=1}^{n} y_{ik}$.

Model-based correspondence analysis

? showed that CCA approximates the solution from maximising the log-likelihood function (1.3) with f_k being Gaussian response functions under conditions 1- 4. Assume that the abundances Y_{ik}, are independently Poisson distributed and that the Poisson parameter can be modelled by means of the species response function $f_k(\cdot;\boldsymbol{\theta})$, with $\boldsymbol{\theta}$ the vector with the parameters of the species response functions. In particular,

$\mathrm{E}(Y_{ik} \mid z_i) = f_k(z_i; \boldsymbol{\theta})$. The Poisson model gives rise to the following likelihood function

$$\text{likelihood}(\boldsymbol{\theta}) = \prod_{k=1}^{K} \prod_{i=1}^{n} \frac{e^{-f_k(z_i;\boldsymbol{\theta})} (f_k(z_i; \boldsymbol{\theta}))^{y_{ik}}}{y_{ik}!}.$$

The corresponding log-likelihood function then equals

$$\text{log-likelihood}(\boldsymbol{\theta}) = \sum_{k=1}^{K} \sum_{i=1}^{n} -f_k(z_i; \boldsymbol{\theta}) + y_{ik} \log f_k(z_i; \boldsymbol{\theta}). \tag{1.3}$$

in which the constant has been left out because it does not depend on $\boldsymbol{\theta}$.

Despite the contributions of CCA to the recognition of ordination analysis in ecology as an analytical tool, its rigid assumptions are controversial. Especially the unit tolerance assumption makes the method often unrealistic in many applications. Fortunately, after the connection between CCA and generalise linear models (GLM) was revealed, attempts have been made to relax the assumptions [**??**].

? proposed an algorithm for direct likelihood maximisation, which allows for the estimation of all parameters simultaneously. The complexity of the approach lies in the fact that the environmental gradient is shared by all K species, and thus all K GLMs must be considered simultaneously by the algorithm. Although Yee's quadratic reduced-rank vectorised GLMs (QRR-VGLM) method [**?**] does the job, it is computationally intensive.

CCA is an example of a *constrained* ordination method, because the analysis of the abundance matrix is constrained by modelling relationships between abundances and environmental variables from a second data source. For this reason, CCA often stands for *Constrained Correspondence Analysis*. We will use the two meanings of CCA interchangeably.

Flexible correspondence analysis

Two years later, **?** came up with a likelihood-based method that is inspired by Fisher's linear discriminant analysis (LDA). The latter method basically consists in projecting multivariate observations $\boldsymbol{x}_i$ onto a vector $\boldsymbol{\alpha}$ such that the resulting scalar scores show maximal between-group sum of squares as compared to the within-group sum of squares. In order to make the step towards an LDA-likelihood-based method for CCA, **?** embedded the data generating mechanism into a probabilistic model, which is described in the next paragraph.

As before, let $z_i = \boldsymbol{\alpha}^T \boldsymbol{x}_i$ denote the environmental scores for a given $\boldsymbol{\alpha}$. Assume that the z_i have been sampled from some distribution with density function $g(\cdot)$. Under this assumption, we use the notation Z_i to stress that it is considered as a random variable, while still using z_i for the observed outcomes. In particular, $Z_i, \ldots, Z_n \sim g(\cdot)$. This encompasses the random sampling of n sampling locations $i = 1, \ldots, n$. At each sampling location the abundances of K species are assumed to be sampled from a Poisson distribution with mean $\mathrm{E}(Y_{ik} \mid z_i) = f_k(z_i)$, $k = 1, \ldots, K$. Next a hypothetical sampling scheme is constructed for species k: at each location i the environmental score z_i, which was sampled from $g(\cdot)$, is replicated y_{ik} times, i.e., y_{ik} is considered as a weight in the probabilistic model. Let $p_k(z)$ denote the density function of the weighted or replicated sample $\{(Z_i; Y_{ik})\}$ in which the Y_{ik} are to be interpreted as the weights. Theorem 1 in **?** states that $p_k(z) \propto g(z) f_k(z)$, which is the product of a density $g(z)$ and a conditional mean from a Poisson model. In a next step they consider the response function $f(z)$, which is referred to as the null model and expresses that the environmental conditions do not affect the species abundances through the scores

$\boldsymbol{z} = \boldsymbol{\alpha}^T \boldsymbol{X}$, i.e., $f(z)$ represents the species response function under the assumption that all species respond in the same way to the environmental conditions. Under this null model, the density function of the weighted sample $\{(Z_i; Y_{ik})\}$ is given by $p(z) \propto g(z)f(z)$. Within this probabilistic sampling framework, a log-likelihood ratio statistic can be defined for contrasting the null model with the model that allows for species-specific response functions,

$$\mathsf{LR}(\boldsymbol{\alpha}) = \log \frac{\prod_{i=1}^{n} \prod_{k=1}^{K} \left[p_k(\boldsymbol{\alpha}^T \boldsymbol{x}_i)\right]^{y_{ik}}}{\prod_{i=1}^{n} \prod_{k=1}^{K} \left[p(\boldsymbol{\alpha}^T \boldsymbol{x}_i)\right]^{y_{ik}}}. \tag{1.4}$$

? argue that finding the $\boldsymbol{\alpha}$ that maximises $\mathsf{LR}(\boldsymbol{\alpha})$ essentially is a generalisation of CCA: the maximiser of $\mathsf{LR}(\boldsymbol{\alpha})$ is the environmental gradient that maximally discriminates the two probabilistic models. A better appreciation of the LR-criterion arrises after substituting the p_k and p with $c_k g(z) f_k(z)$ and $cg(z)f(z)$, respectively, where c and c_k are normalisation constants that do not depend on $\boldsymbol{\alpha}$, resulting in

$$\mathsf{LR}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{k=1}^{K} y_{ik} \log \frac{f_k(\boldsymbol{\alpha}^T \boldsymbol{x}_i)}{f(\boldsymbol{\alpha}^T \boldsymbol{x}_i)} + \text{constant}. \tag{1.5}$$

Hence, the LR-criterion is invariant to the distribution $g(z)$. One can now also interpret the maximiser of the LR-criterion as the environmental gradient that maximally separates the species response functions. The response function $f(z)$ can be derived from the species-specific response functions upon recognising that the density function $p(z)$ from the null model is a mixture distribution: $p(z) = \sum_{k=1}^{K} b_k p_k(z)$, with b_k the relative frequency of species k over all sampling locations. This gives $f(z) = \sum_{k=1}^{K} b_k f_k(z)$, and

(up to a constant term)

$$\text{LR}(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik} \log\left(\frac{f_k(\boldsymbol{\alpha}^T\boldsymbol{x}_i)}{\sum_{j=1}^{K} f_j(\boldsymbol{\alpha}^T\boldsymbol{x}_i)}\right) - \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\log b_k. \tag{1.6}$$

The second environmental gradient can be found by maximising the LR-criterion by substituting the environmental matrix $\boldsymbol{X}$ by the residuals from regressing the environmental variables on z_1, i.e., the first environmental score.

The model-based likelihood method imposes no assumption on f_k; thus the species response function can be of any desirable shape. For this reason the method was names *flexible constrained correspondence analysis*. Maximising (1.6) gives exactly the maximum likelihood solution of $\boldsymbol{\alpha}$ when a conditional multinomial model is considered. The latter is obtained by starting from a Poisson distribution and conditioning on the total abundances at all sampling locations. **?** argued that the multinomial model generates competition among the species.

The equivalence between CCA and LDA had been shown before [**??**]. In particular, under the CCA-assumptions of equal tolerances and Gaussian response functions, the estimate of $\boldsymbol{\alpha}$ is the direction in the p-dimensional environmental variables space along which the species vary most in terms of their optima μ_k $(k = 1, ..., K)$. This characteristic was utilised by **?** as an alternative way for constructing the CCA solution. **?** give the LR criterion and the LDA interpretation for a larger class of models. Finally, **?** concluded that their approach is flexible in the sense that the response functions may even be estimated nonparametrically by realising that the $p(z)$ and $p_k(z)$ density functions can be estimated nonparametrically from the weighted sample $\{(z_i, y_{ik})\}$.

1.2.4 Illustration of ordination analysis and the ordination diagram

We use the hunting spider data set (details are given in Section 1.4.3) to demonstrate three ordination analysis methods (CA, CCA and the model-based likelihood method with the Poisson model). Although there are many types of graphical representations in CA or CCA (see e.g., **??** for more details), in this thesis we will always use the term *ordination diagram*, which can be constructed and interpreted as follows. An ordination diagram is usually a two dimensional plot of which the two axes correspond to the first two environmental scores, i.e., $\boldsymbol{\alpha}_d^T \boldsymbol{X}$ where $d = 1, 2$ refers to the dimension. Each sampling location i is shown as a dot (or site label) with coordinates (z_{i1}, z_{i2}). Each species k is shown as a dot (or species label) which its coordinates are given by the species-specific optima in the first two dimenions: on the x-axis the optima with respect to the response function of the first environmental score, and on the y-axis the optima with respect to the response function of the second environmental score. The graph also shows the loadings of the environmental gradients contained in $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ (presented as arrows). When a species is close to a site, this can be loosely interpreted as that the species is expected to be highly abundant at that particular site. The arrows corresponding to the environmental variables can help in characterising the site in terms of its environment.

Figure 1.4 shows the ordination diagrams for the three different ordination methods. Here are some conclusions:

- From the left and middle panels we can read that species *Pardosa monticola* is abundant at site 1.
- The middle panel (CCA) further suggests that site 1 is covered with moss (Cov-

eMoss), but does not show much fallen twigs and leaves on the ground (FallTwig).

- The middle panel (CCA) shows that sites 8, 21, 20 and 19 lie along the direction which variable *FallTwig* points into. Hence, the percentages coverage of fallen leaves and twigs are high at these sites.

Note that the last conclusion is contradicted by the right panel (Poisson model). The CCA ordination diagram (middle panel of Figure 1.4) and the ordination diagram based on the Poisson model (right panel of Figure 1.4) are different in the sense that the latter does not impose the equal maxima and unit tolerance assumptions to the species response curves and this may result in a different conclusion, as illustrated here. Figure 1.5 is provided to help readers understand the difference in response curves resulting from the rigid assumptions. The left panel of Figure 1.5 shows the species response curves in first dimension under the CCA-assumptions, whereas the right panel of Figure 1.5 gives the response curves obtained from the Poisson model where no such explicit assumptions are imposed. The total mean squared error of the fits of the 12 species response curves for the CCA method in the first dimension is 148.67 whereas the total mean squared error for the more relaxed Poisson method is 52.98. This indicates a overall better model fit of the less restrictive model-based likelihood approach.

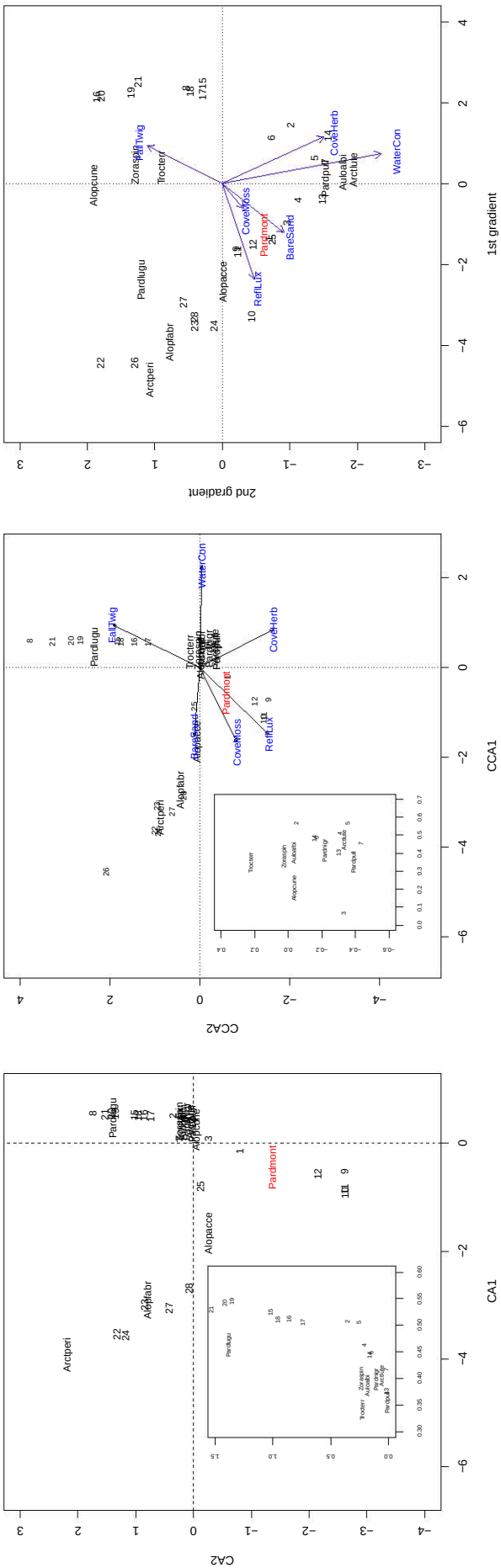


Figure 1.4: Ordination diagrams for the spider data, obtained with CA (left), the CCA (middle) and the Poisson (right) models. Numbers indicate sites and the abbreviation of species name in the legend are given in the caption of Table 1.1. The insets give details of the regions where numbers overlap.

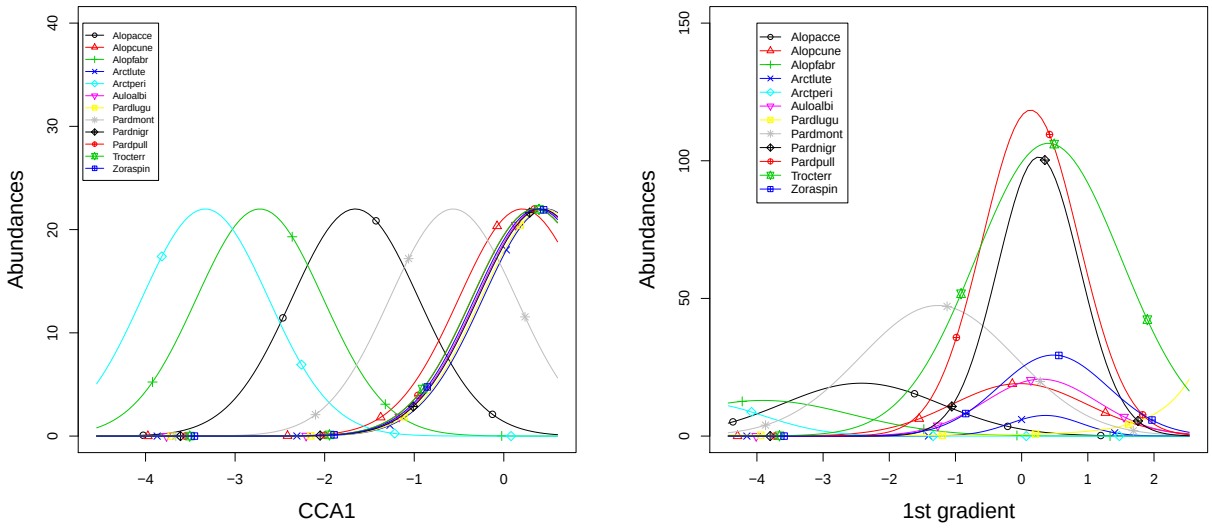


Figure 1.5: The fitted species response curves for the spider data, obtained with CCA (left) and the Poisson model (right). The abbreviations of the species names in the legend are given in the caption of Table 1.1.

1.3 Species diversity indices

Species diversity refers to the number of different species in a community and their abundances. It consists of two components: species richness and species evenness. The former is the number of different species in a given area, while the evenness is related to the similarity of the abundances. For example, suppose that in a given region five different species live, and suppose that the abundances of all five species are equal. Then the richness is equal to five and we say that it is a completely even community. A very uneven community arises when, for example, the relative abundances are 96% for one species, and 1% for each of the other four species. The evenness index ranges from about 0, when only few species are highly abundant, to 1, when the relative abundances of the species are nearly equal. The richness index lacks information on the composition of the community. Therefore evenness is often considered

as a more informative measure. Several mathematical formulations have been brought up to quantify biodiversity [**???**], among which the Gini index [**?**] is appreciated due to its generality.

? argued that together with species diversity the scale must be integrated into the numerical analysis. He suggested defining species diversity at two scales: a local scale and a larger regional scale. The former often refers to a single location or sampling site, and the latter to a larger area or landscape. The species diversity at the local scale is named the *alpha diversity* (α-diversity), and at the regional scale it is named the *gamma diversity* (γ-diversity). The ratio of γ-diversity over α-diversity is referred to as the *beta diversity* (β-diversity).

In this thesis we will focus on the α-diversity, for which several indices have been proposed. We limit the discussion here to three popular indices: the Simpson, Shannon and the Gini indices. Since all indices refer to the α-diversity at a single sampling location, we simplify the notation by dropping the location index i in Y_{ik}. Many indices are based on the relative abundances of the species present at the sampling location. We use the notation p_k to refer to the relative abundance of species $k = 1, \ldots, K$. For the Simpson, Shannon and Hill indices, it does not matter whether or not the p_k are defined in terms of the expected abundances or the observed abundances. If the latter is used in the formulae following, then the indices are to be considered as consistent estimators of the indices defined at population level.

The *Simpson index* is given by [**?**]

$$S = \sum_{k=1}^{K} p_k^2. \tag{1.7}$$

Note that S is bounded between $1/K$ and 1. It has a probabilistic interpretation: it is the probability that two individuals randomly selected from the same location, belong to the same species. A perfect even community will thus get a Simpson index of $1/K$ and a maximally uneven community receives a Simpson index of 1. Hence, the larger S, the more uneven the community. The same information is often reported as the *inverse Simpson index*, $1/S$, or the *Gini-Simpson index*, $1 - S$.

Another popular diversity measure is the *Shannon index* [**?**]. It is calculated as follows:

$$H = -\sum_{k=1}^{K} p_k \ln p_k. \tag{1.8}$$

The index H ranges from 1 (in the limit) for a community of infinite diversity to $\ln K$ for a complete even community. It expresses the uncertainty in predicting to which species a randomly selected individual belongs [**?**]. Hence, the larger H, the more even the community. Note that the Shannon index is directly related to the Shannon entropy, which is sometimes considered as a measure for chaos.

? united several diversity indices. He defined the diversity number of order q as

$$H_q = -\left(\sum_{k=1}^{K} p_k^q\right)^{1/(1-q)}.$$

When $q = 0$, this reduces to the richness. For $q = 2$, the inverse Simpson index is retrieved, and, in the limit as $q \to 1$, H_q converges to the Shannon index.

The Simpson index is claimed to be more ecologically meaningful [**?**] as it gives an intuitive interpretation in probability. **?** argued that the Simpson index is less sensitive to the sample size.

The Gini index, also known as the Gini coefficient, is a measure of inequality that has been frequently employed by economists for quantifying the inequality among income or wealth. It is conventional to introduce the Gini index through the Lorenz curve, which was developed by Max O. Lorenz in 1905 [**?**].

For defining the Gini coefficient we need the distribution function of the marginal distribution of the abundances at a sampling location. In the previous parts of this chapter we used Y_k to denote the abundance of species $k = 1, \ldots, K$, and we sometimes assumed a particular distribution for Y_k, say $F_k(y)$ (e.g., a Poisson distribution). Using a different notation, we could write $Y \mid$ species $k \sim F_k$, which also states the conditional distribution of abundance Y, given species k. The marginal distribution of the abundance Y is then simply obtained by marginalising over all K species at a particular sampling location. The marginal distribution function is denoted by F. In other words: the marginal distribution $F(y)$ of Y gives the distribution of abundances of randomly selected species. We will also need the inverse distribution function (quantile function), denoted by F^{-1}.

We first define the Lorenz curve. Let μ denote the mean of Y. The Lorenz curve is then given by

$$L(p) = \frac{1}{\mu} \int_0^p F^{-1}(u) du.$$

Hence, $L(p)$ is the expected abundance of the $100 \times p\%$ least abundant species, divided by the mean abundance of all species at the sampling location considered (i.e., $L(p)$ gives the expected relative abundance of the $100 \times p\%$ least abundant species).

Figure 1.6 shows a typical Lorenz curve. The diagonal line in Figure 1.6 represents a perfect even situation. The Gini index is defined as the shaded area A. Note that

$A + B = 0.5$, the Gini index is thus given by

$$G = A = \frac{1 - 2B}{2},$$

where B is the area under the Lorenz curve,

$$B = \int_0^1 L(p)dp.$$

In economics, the Lorenz curve is used for quantifying the wealth inequality in a population. If Y is the income of a household, then $L(p)$ gives the total income of the $100 \times p\%$ poorest households, rescaled by dividing by the mean income μ.

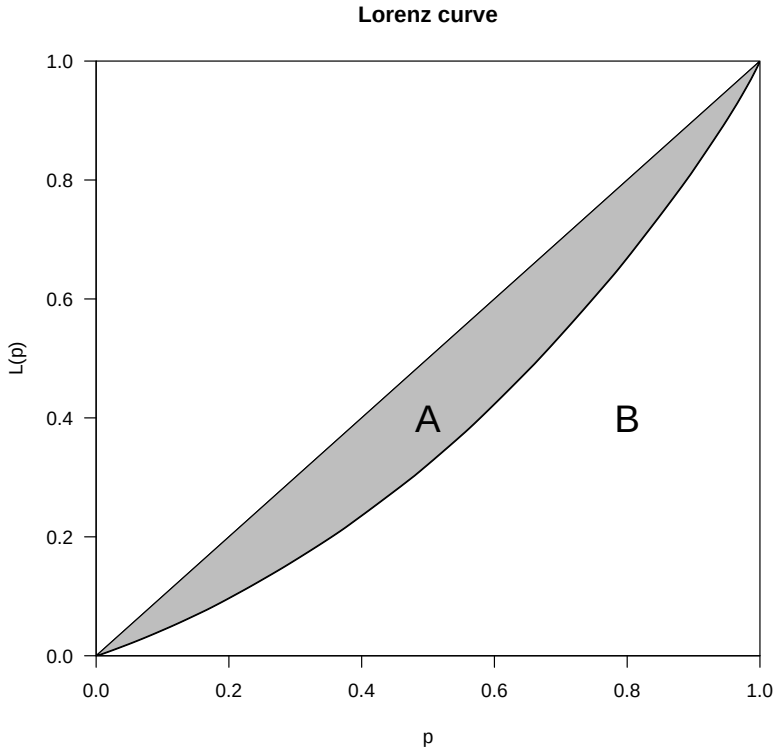


Figure 1.6: Example of a theoretical Lorenz curve. p is the cumulative proportion of number of species and $L(p)$ is the corresponding marginal cumulative proportion of abundance. The Gini index is defined as the the area of A.

Several papers have been published to compare behaviours of the different diveristy indices [**???**], but no clear conclusion was made regarding the outperformance of a particular index. Nevertheless, the properties of the Gini index have attracted many studies due to its comprehensible mathematical interpretation. For a particular sampling location i, a nonparametric estimate of the Gini index given by **?** is

$$2\hat{G}_i = \frac{1}{\bar{Y}_i} \times \frac{\sum_{k=1}^{K} \sum_{l=1}^{K} |Y_{ik} - Y_{il}|}{K(K-1)}, \tag{1.9}$$

where $\bar{Y}_i$ is $\sum_{k=1}^{K} y_{ik}/K$. The variance of $\hat{G}$ is discussed by **?**; this enables comparing the species diversity across locations. In Chapter 4 we focus on the estimation of G and on modelling of G as a function of environmental variables. With increasing desire of learning the effect of a large number of chemical and physical environmental factors on species diversity, researchers attempt to bring up appropriate prediction models for the Gini index. Few papers [**???**] were published on forecasting diversity indices from variables of habitat characteristics, among which the multiple linear regression (MLR) approach is currently most widely used. The theory is rather simple: the MLR uses a nonparametric diversity index estimate as the response variable and a set of q environmental variables of interest as the regressors in a linear model. This approach actually involves two separate steps: in step one the site-specific diversity indices are computed solely based on the abundance information and then in step two a linear regression is implemented. The two-step structure may result in incorrect statistical inferences because the variances of the estimated diversity indices may not be equal among sampling locations, which violates an assumption of the linear regression method.

1.4 Data sets

In this section data sets used for the demonstration our new methods are introduced. Two different types of data are used: (1) two datasets from classical macroecological abundance studies on animals; (2) two datasets from microecology. Because the technology for quantifying the microbial abundances in the two latter data sets is rather recent, and because later we need to refer to technological aspects to explain certain issues in the data, we devote a section to the description of the sequencing technology used to generate the data. This technology is part of *metagenomics*, which is the topic of Section 1.4.1. Sections 1.4.2 and 1.4.3 give introductions to the metagenomics and animal abundance data sets, respectively.

1.4.1 Metagenomics and 16S rRNA sequencing

Biology and technology

Different from the classical isolation and cultivation-dependent microbial ecology studies, in metagenomics the microbial organisms are directly studied in their environment, i.e., without the need for laboratory cultivation of individual species.

In very general terms, metagenomics is based on the analysis of all genomic material present in a sample taken from a sampling location (either environmental, human, . . .). The exact definition of metagenomics is still under debate in the scientific community. We will use the term in a very strict form: identification of all known microorganisms in a sample, and we particularly focus on the sequencing of the 16S ribosomal RNA (16S rRNA) genes of microorganisms. The 16S ribosomal RNA is a part of the ribosomes of

prokaryotes, which are single cell organisms to which bacteria belong. Ribosomes are complex molecular structures in cells that play an important role in the translation of mRNA into a polypeptide chain. The structure is composed of several subunits which consists of proteins and of a few RNA molecules which are referred to as *ribosomal RNA* (rRNA). In contrast to mRNA, the rRNA molecules themselves are not translated into proteins; they play a role in the functionality of the ribosomes. However, just like mRNA the rRNA is a transcript product of genomic DNA and it can therefore be sequenced using sequencing technologies (see further). These 16S rRNA genes are very well suited for the identification of bacterial species, because it contains regions that vary between species, but are highly conserved within species. Several reference databases, which connect the 16S rRNA gene sequences to bacterial species, are available. Another advantage in the use of 16S rRNA is that these interesting regions can be easily amplified (necessary preprocessing step in most sequencing technologies) because many universal PRC primers are available for the highly conserved regions. A final advantage is that the method allows for the identification of species based on only a very specific genomic DNA region so that many samples can be sequenced simultaneously, and the cost is strongly reduced to, e.g., whole genome shotgun sequencing methods for species identification. The latter basically consists in the fragmentation of the whole DNA genome into small fragments that are subsequently amplified and sequenced. The method allows to sequence the whole genome, and its has the advantage of giving less species identification errors, but it is more expensive and the genome assembly step is time consuming and also error prone. In this thesis we use two datasets obtained through 16S rRNA sequencing.

A typical work flow for 16S rRNA metagenomic data generation is shown in Figure 1.7.

It starts with sample collection. The collected samples are first purified and DNA is extracted. However, the different DNA fragments are mixed together and they need to be separated for sequencing. The step of separating the sequences is part of the process called library construction. The target region of 16S rRNA is amplified throng Polymerase Chain Reactions (PCR). The amplified 16S rRNA fragments are then sequenced using a massive parallel sequencing technique. There exist several sequencing platforms; we refer to **?** for an comparison of some of the major platforms. The output of the sequencing device is a large set of reads. Each read is a sequence of nucleotides originating from a DNA fragment. Because the sequencing starts from PCR amplified DNA fragments, each original DNA fragment may be sequenced multiple times. The average number of reads that cover a nucleotide is referred to as the *sequencing depth* or *coverage*. It is expected that the larger the coverage, the less error prone the subsequent statistical analysis. The total number of reads produced by the sequencing experiment of one sample, is known as the *library size* of the sample. Once the reads are available, the wet-lab handling is over and the species identification process continuous with data-processing steps (bioinformatics).

Next, the resulting sequenced reads are clustered into groups of closely related sequences; this step is called binning. A binning method can be based on the similarities among the sequences or the similarity of a sequence to known references [**?**]. The reads can be binned according to different levels of the similarity. A cut-off of 97% similarity is often applied to obtain Operational Taxonomic Unit (OTU) level, which is a pragmatic proxy for the microbial “species” taxonomic levels. Reference databases are available for OTU (or species) identification. The sequencing technology does not only allow for the identification of the OTUs present in a sample, but the number of

reads mapped to an OTU is also considered as a proxy for the abundance of the OTU in the sample. Hence, the data can be represented as an abundance matrix as shown in Figure 1.1. Starting from the OTU classification, data can also be represented at higher taxonomic ranks, using, e.g., the bacterial phyla classification.

Two important characteristics of the abundance data:

- The rare species are prone to be undetected in the community due to sequencing error or insufficient sequencing depth or coverage [**?**].
- The total number of sequences or reads (library size) varies form sample to sample due to different wet-lab handling and differences in PCR efficiency during the amplification step. This variation can not be controlled nor is it associated with true abundance [**?**]. As a consequence, more rare OTUs will be observed in the samples with larger library size.

As a consequence, the probability of observing zero abundance also depends on the library size and insufficient library size may lead to zero inflation [**?**].

Examples of metagenomics studies

With the advancement of genomic sequencing methods and the drop of the cost, the number of metagenomic projects is drastically increasing. More and more significant discoveries are made through metagenomics studies. For example, the Earth Microbiome Project is set to attempt to characterise the functional diversity of global taxonomic for the benefit of the planet and the human being. Another example is the Human Microbiome Project (HMP) which is set to learn the correlation between changes in the

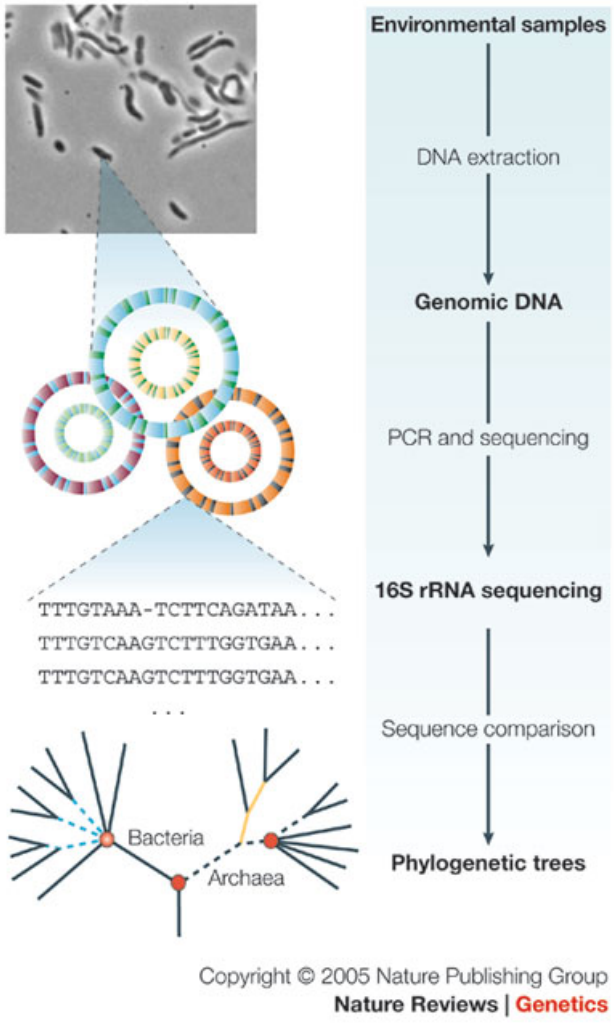


Figure 1.7: The work flow of 16S rRNA sequencing [**?**].

microbial community found at multiple human body sites and human health [**?**].

The downstream statistical analysis includes exploratory data analysis, e.g., Principle Component Analysis for revealing the relationships between taxa, and CCA for studying taxa-environment relationships.

Many human microbiome studies aim at the detection of bacterial species that show different (relative) abundances between groups of people, e.g. between healthy and obesity subjects. In environmental metagenomics studies, the question may, e.g., relate to differential abundance of species between two regions with different climate conditions. These questions are typically answered by means of large scale statistical hypothesis testing. This usually involves testing for differential abundance at individual species level, and subsequently correcting for multiple testing so as to control the false discovery rate at a desired level. Instead of performing a statistical analysis for each species separately, the research may also focus on the community structure as summarised into a biodiversity index. See **?**, **?** and **?** for examples.

1.4.2 Metagenomics data sets

Antarctic lakes data

The data set is from a limnology study described by **?** and the sequencing platform used for data generation is Roche 454. The study objective was to learn how microbial communities in Antarctic lakes respond to the environmental changes in the lakes. A CCA seems an appropriate data exploration method that serves this research question. The researchers collected 45 water samples from lakes in ice-free regions along the

east Antarctic coastline. Within each water sample 13 physical and chemical characteristic were measured (environmental variables), including the depth at which the sample was taken, pH, conductivity, the concentrations of major ions (Na^{2+}, NH_4^+, SO_4^{2-}, K^+, Mg^{2+}, Ca^{2+} and Cl^-), silicate, the total organic carbon (TOC) and the dissolved organic carbon (DOC). More than 500 microbial species were identified and their relative abundances quantified using 16S rRNA sequencing on a Roche 454 device.

Rare species are common in the data set: only 199 species appear in no less than 3 sampling locations. Figure 1.8 gives a graphical presentation of the proportion of zero abundance in the data. Table B.1 in Appendix B.1 gives a summary of the relative abundances of each species and Table B.2 in Appendix B.1 shows an overview of the environmental data.

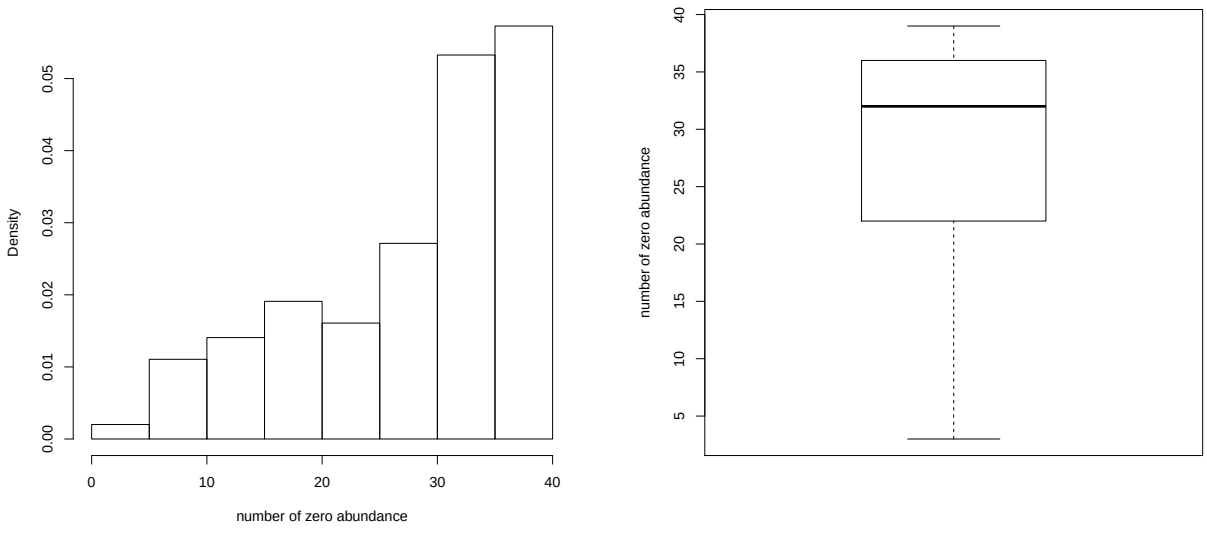


Figure 1.8: Histogram of the number of zero abundances of each species (left) and boxplot of the number of zero abundances (right) in the Antarctic lakes data.

Human infant gut microbiome data

The infant gut microbiome data set originates from a longitudinal study of **?**. The original research objective was to identify the link between the development of human infant gut microbiome and the development of type 1 diabetes (T1D). Thirty-three infants who were predisposed to develop T1D were signed up to the study and followed from birth until 3 years of age. Stool samples of each infant were taken on a regular basis to assess the composition of the gut microbiome. In total 777 samples were sequenced by 16S rRNA sequencing on the Illumina MiSeq V2 platform [**?**]. This resulted in 2239 unique OTUs identified. Additionally the dietary intakes of each infant were kept track of. Sixteen diet-related variables are recorded including, for example, whether the infant is fed with solid food at the time the sample was collected. In this thesis, for the purpose of demonstration of the developed methods, the analyses are applied to the family level of the microbes. A total of 220 OTUs are left out of the analysis since unidentified. In total the abundance matrix contains for each of the 777 samples the abundances of 48 distinct microbial families. Figure 1.9 shows the proportion of zero abundances in the data set: the median of the number of zero abundances is close to 400. Table 5.1 in Chapter 5 gives an overview of the 16 dietary variables.

Chapter 5 is completely devoted to this data set. One of our new CCA methods will be applied, as well as our new Gini index regression model.

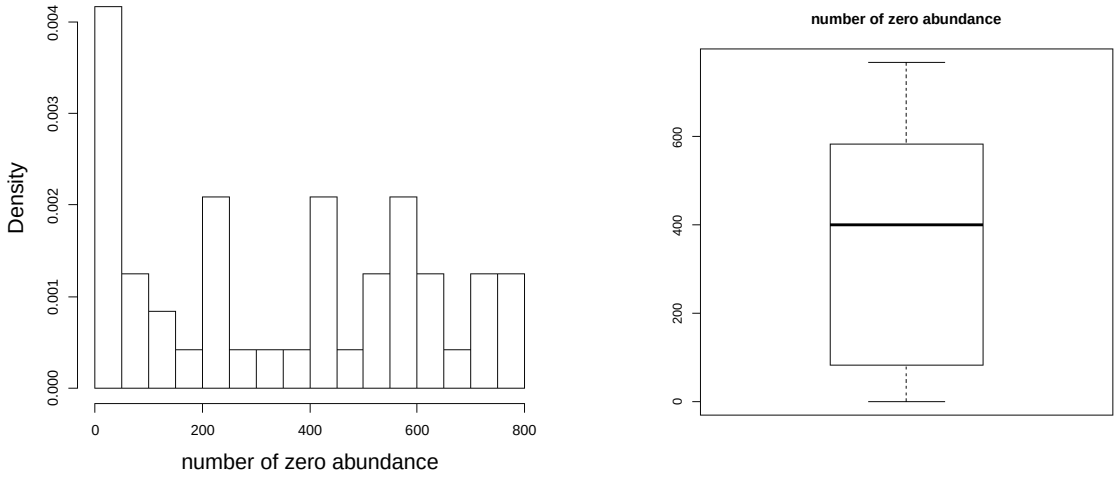


Figure 1.9: Histogram of the zero abundances of each species in the data set (left) and boxplot of the zero abundances (right) of the Human Infant Gut Microbiome Data.

1.4.3 Animal abundance studies

Hunting spider data

The data set originally comes from an experiment conducted by **?** in a Dutch dune area. This data have been used by **?** and **?** for the demonstration of their ordination methods for revealing the changes in spider community in relation to the environmental changes. The abundance data of 12 different spider species were collected at 28 locations over a period of 60 weeks. Six measurements of the habitat, including percentage of dry mass, percentage over of bare sand, percentage over of fallen leaves and twigs, percentage cover of the moss layer, percentage cover of the herb layer and reflection of the soil surface with cloudless sky were recorded. Tables B.3 and B.4 in Appendix B.2 show a summary of the abundance data set and environmental data set. Figure 1.10 show the abundance distribution of each spider species (left panel) and the distribution of zero abundance (right panel).

We will use this data set to illustrate our CCA methods.

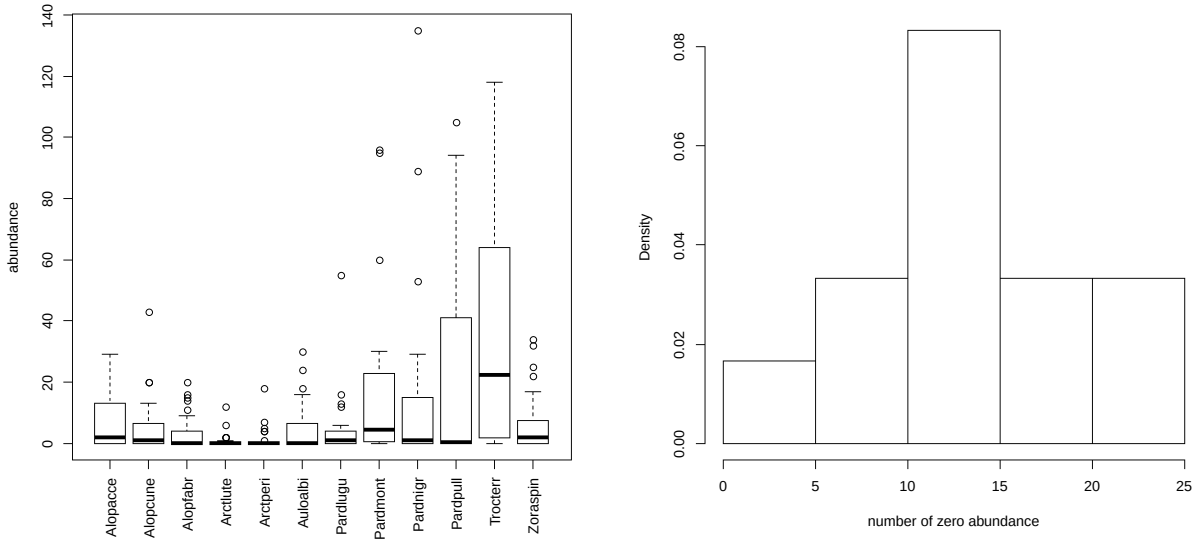


Figure 1.10: Boxplot of the abundances of 12 spider species (left) and histogram of the number of zero abundances of each species (right) in the Dutch dune hunting spider data.

Oribatid mite data

? conducted a field survey on the southern shore of a small Laurentide lake on the station de Biologie des Laurentides, Canada. The study objective is to learn the environmental and spatial influence on the mite community variation. The sampling area is a 10×2.6m transect vegetation mat surrounding the lake and it ranges from the forest border to the water front. The sampling area is geographically subdivided into 7 regions by the dominating 7 different vegetation types. The different vegetation types are called 'substratumtypes' in the original paper. Within each region, several representative spots of 2×2 cm are selected. The resulting data contain the abundance matrix and the environmental data matrix. The oribatid mite abundance matrix contains the counts of 35 adult orbited mites at 70 different spots. The environmental data

matrix contains the density of the substratum, the water content, the substratum type (7 classes) and coverage density of the shrub. Tables B.5 and B.6 in Appendix A.3 show summary statistics of the abundance data set and a detailed description of the environmental data set.

The data are used to illustrate our CCA methods.

1.5 Challenges

Large numbers of zero abundances is a common phenomenon in community ecology, particularly when many species are observed, which particularly happens when high-throughput sequencing technologies are used. A zero abundance may be a *true zero* which represents the absence of a species under particular environmental conditions at the sampling sites. Zero abundances may also be observed due to sampling related processes. For example, a species may be present at a sampling site, but because of the randomness associated with the sampling process, no individual of the species ends up in the sample. Many sampling processes result in a Poisson distribution for the abundances of a given species at a given sampling site: the Poisson distribution describes the abundances observed if the same site were sampled repeatedly. This type of variability is referred to as technical variability. The Poisson distribution is also appropriate for describing the distribution of the read counts for a given OTU in technical replicates, i.e. distribution of read counts of a given OTU when the same biological sample is repeatedly analysed (sequenced). However, in microbial community ecology, metagenomics high-throughput sequencing shows small sensitivity for rare species, resulting in too many observed zeroes as compared to expected. Furthermore, more

zeroes caused by the insensitive measurement technologies are expected with increasing numbers of species. These zeroes correspond to species that are present in the sample, and these zeroes are therefore considered as *false zeroes* and they need to be modelled by distributions that allow for such zero-valued observations. Figure 1.11 shows histograms of four randomly selected spider species. It is obvious that for each case the frequency of zero abundance is too high to be well modelled by a Poisson distribution. As neither CCA nor its model-based counterpart take the zero inflation into account, the results from these approaches are potentially inaccurate. Part of this thesis is devoted to the development of ordination analysis methods that explicitly take zero inflation into account so as to result in more accurate results, in the sense that the final conclusions are less error-prone due to these zeroes.

The underlying ecological assumption of both the model-based likelihood method and the eigenanalysis-based CCA is the fundamental niche theory which describes the way a species abundance responds to the environmental conditions. The fundamental niche theory says that a species will be abundant under the environmental conditions that it favours, and its abundance will decease as the the environmental conditions move away from the optimal conditions. This relation is mathematically described by a unimodal response curve (Figure 1.2), which is often taken to be bell-shaped. The expression in Equation (1.1) will always give a bell-shaped curve, but when using the GLM procedure, which is part of the likelihood-based methods, the bell-shaped feature can only be guaranteed if the coefficient of the second order polynomial (β_{2k} in Equation (1.2)) is negative. CCA will always result in bell-shaped response curves because of the unit tolerance assumption. By equating the parameters in Equation (1.1) to the parameters in Equation (1.2), the unit tolerance leads β_{2k} to be -0.5 for all species.

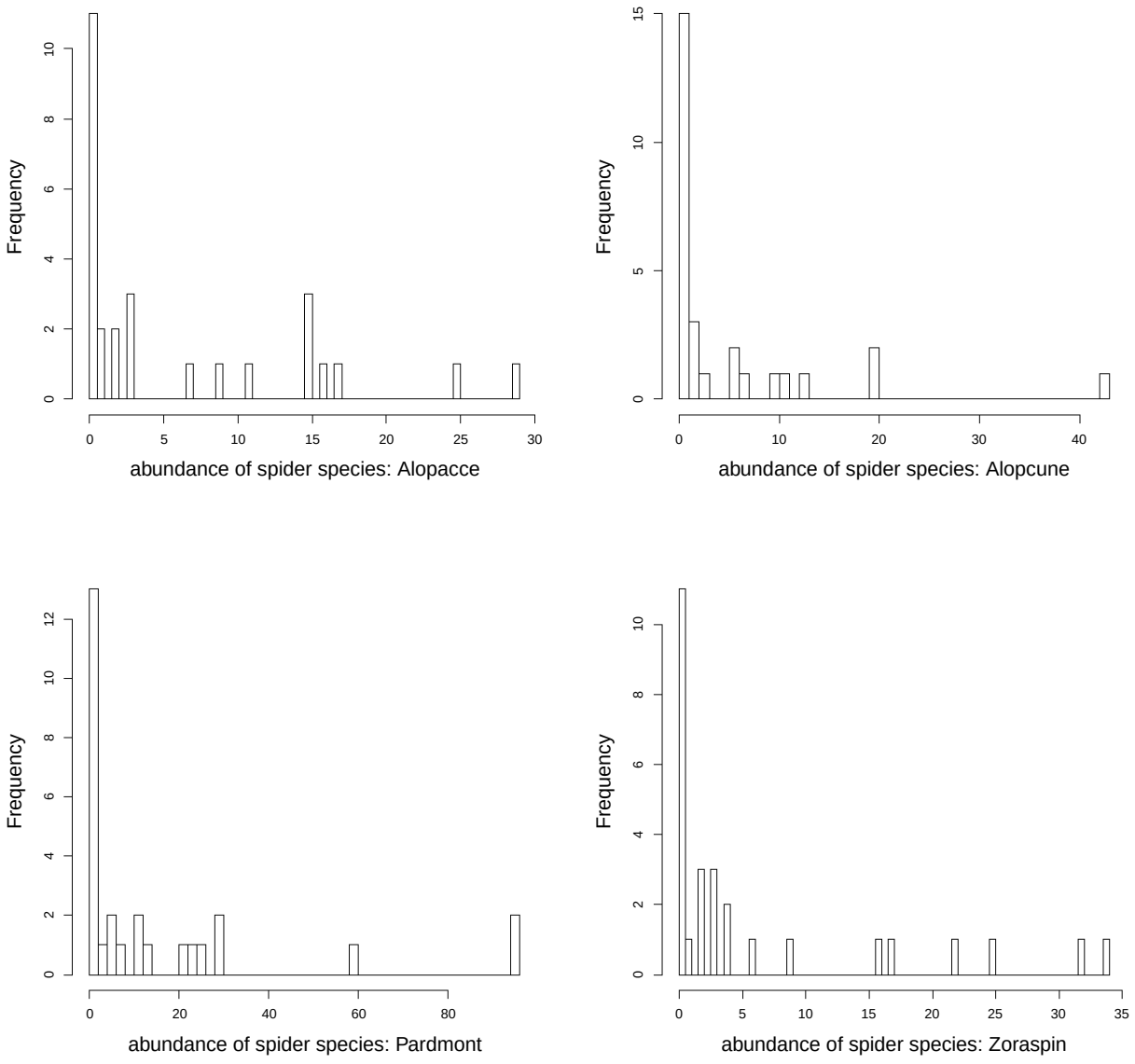


Figure 1.11: Histograms of the abundances of 4 randomly selected spider species.

However, the unit tolerance assumption is relaxed in the model-based likelihood approach, and hence bell-shaped response curves are no longer guaranteed. In fact, this can be seen in the right panel of Figure 1.5 where the response curve of species *Trochosa terricola* is not bell-shaped. As a consequence, the information presented in the ordination diagram could be misleading because the position of *Trochosa terricola* in the ordination digram corresponds to a minimum rather than a maximum (optimum). To our knowledge, most users are unaware of the risk associated with interpreting these ordination graphs. Being motivated by the problem, in this thesis we will explore ways to adjust the model-based likelihood method so as to ensure bell-shaped response functions. This may consequently leas to less error-prone interpretations of ordination plots.

Consider again the spider data. Previously we have mentioned that the classical two-step approach for modelling the diversity index may give wrong statistical inference because the variance heterogeneity of the estimated Gini indices is ignored. We can see this from Table 1.3 which shows the estimates of the Gini indices at 28 sampling locations and their variance (estimated following **?**). In this thesis we propose a semi-parametric regression method for modelling the Gini index as a function of environmental variables. Our method is a one-step method and deals correctly with the variance heterogeneity.

1.6 Objectives and outline

The objectives of this dissertation are the development of

Table 1.3: The estimated Gini indices ($\hat{G}$) of 28 sampling locations of the spider data set and the estimated variances $\hat{\text{var}}_{\hat{G}}$.

Site	$\hat{G}$	$\hat{\text{var}}_{\hat{G}}$
1	0.6310	0.4608
2	0.7073	0.2520
3	0.5857	0.3247
4	0.6639	0.0816
5	0.7006	0.1860
6	0.6229	0.1251
7	0.6376	0.0672
8	0.7455	1.2992
9	0.8333	2.3990
10	0.8673	2.9656
11	0.8550	2.8342
12	0.7829	1.6171
13	0.5809	0.3718
14	0.7078	0.0420
15	0.8578	0.2555
16	0.7965	0.2152
17	0.8796	0.1604
18	0.8790	0.2795
19	0.8098	0.5450
20	0.8184	0.6873
21	0.7799	0.8811
22	0.7778	6.7641
23	0.7083	4.7684
24	0.7885	7.1007
25	0.5423	0.8217
26	0.8274	5.8690
27	0.8352	14.3386
28	0.7588	6.6729

- ordination analysis methods that are robust to zero inflation.
- model-based ordination analysis methods that guarantee biologically meaningful result by enforcing bell-shaped species response functions.
- a semiparametric regression model for modelling the Gini index as a function of covariates.

In Chapter 2 we propose a constrained ordination method that can cope with excess zeroes. Our method relies on the theory of the likelihood-based ordination method of **?**.

We also propose a constrained ordination method that focuses on presence-absence of species, rather than on their abundances. All methods are demonstrated on the hunting spider data set. The work presented in Chapter 2 is published in **?**.

Flexible constrained ordination methods are based on the GLM procedure, but they do not guarantee that the species response functions are bell-shaped, which is a crucial characteristic of the ecological niche theory. Results from ordination methods are often graphically displayed and researchers interpret these graphs as if all species have bell-shaped response functions (i.e. they assume that all species indicated in the plot represent their optimal environmental conditions, as implied the bell-shape). Hence, researchers are seriously at risk of misinterpreting the ordination graphs. In Chapter 3 we extent the flexible constrained ordination method of **?** so that the final solution is enriched with bell-shaped species response functions. This is accomplished by using a penalised maximum likelihood approach in combination with cross validation to select an appropriate degree of penalisation that gives a good balance between the goodness-of-fit of the model and the number of bell-shaped response functions. The method is applied both to absolute abundance data as to absence-presence data. This chapter is published as **?**.

In Chapter 4 we focus on modelling species diversity as a function of environmental measurements. Species diversity is here described through the Gini index, which is a widely used quantity of α-diversity. In this chapter we first reveal the relation between the Gini index and the first two L-moments of the marginal distribution of abundance. Subsequently we propose a new semiparametric model framework to model the second L-moment. This L2-regression model is combined with a regression model for the mean (equivalent to the L1-moment) to result in a semiparametric model for the Gini

index. The asymptotic properties of the model parameters are also established. Our modelling approach implicitly accounts for the heterogeneity of the variance of the Gini index estimator, which is ignored in the classical two-stage regression approach. The work presented in Chapter 4 is new and not published yet. Throughout Chapter 4 we made a few assumption to simplify the problem. Future work will have to focus on relaxing these assumptions. In the discussion (Chapter 6) we give more details on future work.

Over the years I was working on my PhD research, microbiome studies became more and more popular. Many of the issues to which I tried to formulate solutions are also seen in microbiome studies, particularly the presence of an inflated number of zeroes, and the need to model α-diversity as a function of covariates. Therefore, in Chapter 5 we demonstrate the methods of Chapters 2 and 4 by applying them to human gut microbiome data from **?**.

In Chapter 6 some conclusion and further research perspectives are given.

Chapter 2

Constrained Ordination Analysis in the Presence of Zero Inflation

Summary: Constrained ordination analysis, with canonical correspondence analysis (CCA) as its best known method, is a class of popular techniques for analysing species abundance studies in ecology. These methods rely on distributional assumptions on the conditional abundance distributions. For abundance observations, the Poisson and the negative binomial distributions are the most frequently considered distributions. However, many large abundance studies result in many zero abundances. This may happen because of several reasons. To name one, in microbial community ecology the abundances of a very large number of species are nowadays often obtained by means of sequencing the pooled DNA sample. Due to the small sensitivity for rare species, too many observed zeroes are to be expected. Moreover, more zeroes are expected with increasing number of species. We propose a constrained ordination method based on zero-altered count distributions (e.g. zero-inflated Poisson, hurdle

models, . . .). We show how the parameters and the environmental gradients can be estimated. In simulation studies we examine the behavior of the estimators, and we apply the method to a real data set. We conclude that in the presence of zero-inflation our methods give better results than the Poisson based approaches.

Key words: correspondence analysis; dimension reduction; ecology; hurdle model; zero inflated Poisson

2.1 Introduction

Constrained or Canonical Correspondence Analysis (CCA) is a widely used technique among environmental ecologists for studying the abundances of species under different environmental conditions. It was first proposed by **?** as an extension of the correspondence analysis method for exploring the dependence structure in a two-way contingency table or a species-by-sample abundance table (see, for example, **?** for a recent account). CCA allows for relating the differences between species abundances to scores that are defined as linear combinations of the environmental conditions of the sampling locations. When the species-specific response functions (i.e. the mean abundance as a function of the environmental score) are described by a bell-shaped function (Gaussian) and the abundances are assumed to follow a Poisson distribution, a Constrained Gaussian Ordination (CGO) model arises. **?** showed that his eigenvalue-based CCA can be considered as an approximation to the maximum likelihood solution of a CGO model under mild distributional assumptions. The approximation was studied in an empirical simulation study by **?**; they concluded that the approximation is good when the species-specific tolerances are more or less equal (see later). While

the eigenvalue-approach of CCA has the advantage of being computationally more efficient, the model-based likelihood method has the advantage of being more flexible as it allows for replacing the Poisson and Gaussian components with other functional forms, numerical instabilities notwithstanding. We will require the latter approach for the developments of our methods, but CCA will be included in all simulation studies as it is still the standard method for many ecologists.

In many applications, and particularly in large studies, too many species show zero abundances at too many sites for a Poisson distribution to be appropriate. **??** proposed a maximum likelihood approach to treat the large proportion of zero entries in ecological data sets, but their method cannot be fit directly in the CGO setting. In this paper we study constrained ordination analysis methods that allow for zero inflation. In particular, we replace the Poisson distribution with several *zero-altered distributions* (ZAD). In general we consider two families of ZADs: zero-inflated distributions and hurdle distributions. **?** and **?** advocated such models for analysing count data of rare species, but they did not cover constrained ordination related methods. A more detailed motivation to our methods is postponed to Section 2.3.1.

In Section 2.2 the basic CGO model is presented. The new extensions to this classical model are introduced in Section 2.3. The estimation of the environmental gradient is the topic of Section 2.4. The new method is evaluated in a simulation study, which is the topic of Section 2.5. The new method is further adapted to model the absence/presence type data; this is discussed and illustrated in Section 2.6. We conclude with a brief discussion in Section 2.7.

2.2 The Poisson Model

Let $\boldsymbol{Y}$ denote the $n \times s$ abundance matrix with y_{ik} the abundance of species k ($k = 1, \ldots, s$) at site i ($i = 1, \ldots, n$). Sometimes we will use the notation Y_{ik} instead of y_{ik} to make the distinction between the random variable and the observed abundances. The $n \times p$ matrix $\boldsymbol{X}$ contains the observations on the environmental variables: x_{ij} is the observation on the jth ($j = 1, \ldots, p$) variable at site i. An environmental score at site i is defined as a linear combination of the p environmental variables. The score on the mth environmental gradient at site i is denoted by $z_{mi} = \boldsymbol{\alpha}_m^T \boldsymbol{x}_i$, where $\boldsymbol{x}_i^T$ is the ith row of $\boldsymbol{X}$, and $\boldsymbol{\alpha}_m$ is the vector that determines the mth environmental gradient ($m = 1, \ldots, p$). Later we come back to how the $\boldsymbol{\alpha}_m$ are specified. Because we will first focus on the first environmental gradient we will drop the index m. All models in this section are defined in terms of the one-dimensional covariate z. The estimation of $\boldsymbol{\alpha}$ is discussed in Section 2.4.

A fundamental principle of constrained ordination is that the distribution of the abundance of a species at a particular site depends on the environmental conditions at that site. This can be formulated as a generalised linear model (GLM), i.e. the model has three components: (1) a distribution; (2) a linear predictor; and (3) a link function that relates the linear predictor to a parameter of the distribution. We give the description of the classical Poisson CGO model along the lines of **?** and **?**.

In many situations it is reasonable to assume that the abundances can be described by a Poisson distribution and that the mean of the Poisson distribution depends on the environmental variables through the environmental scores. In particular, let ($i =$

$1, \ldots, n$, $k = 1, \ldots, s$)

$$Y_{ik}|\boldsymbol{x}_i \sim Y_{ik}|z_i \sim \text{Poisson}(\lambda_{ik}) \tag{2.1}$$

with $\lambda_{ik} = \mathrm{E}\left(Y_{ik}|\boldsymbol{x}_{ik}\right) = f_k(z_i)$ in which f_k is referred to as the *response function* of species k.

In ecology it is often believed that the response function should be unimodal, which corresponds to a species that has the largest abundance at sites with optimal environmental conditions; the more the environmental conditions deviate from the optimal point, the smaller the abundance is expected. A convenient and popular choice for the response function is the density of a normal distribution (up to a scaling constant). This results in

$$\log f_k(z_i) = a_k - \frac{(z_i - \mu_k)^2}{2t_k^2}, \tag{2.2}$$

where a_k is the maximum log-mean abundance of species k which is reached when the environmental score z_i equals the *optimum* μ_k. The parameter t_k is referred to as the *tolerance*: the larger t_k the larger the tolerance of species k to deviations from the optimum environmental condition.

If $\boldsymbol{\alpha}$ is known, the assumptions describe for each species k a log-linear Poisson regression model with a quadratic effect of the environmental gradient z. For a given $\boldsymbol{\alpha}$, the parameters a_k, μ_k and t_k can be estimated using maximum likelihood. However, the vector $\boldsymbol{\alpha}$ should be estimated too. One approach consists in first estimating the scores z_i and subsequently regressing them on the environmental variables $\boldsymbol{x}_i$. A general algorithm for direct likelihood maximisation, which estimates all parameters simultaneously, has been proposed by **?**. The complexity lies in the fact that $\boldsymbol{\alpha}$ is shared by all s species, and thus all s GLMs must be considered simultaneously by

the algorithm. Although Yee's quadratic reduced-rank vectorised GLMs (QRR-VGLM) method does the job, it is computationally intensive. **?** proposed an alternative estimation method which has computational advantages. More details will be given in Section 2.4. Finally, we note that the traditional CCA solution of **?** is an approximation of the maximum likelihood solution under the additional assumption that all tolerances coincide, i.e. $t_1 = \cdots = t_k = 1$.

2.3 Zero-Altered Models

2.3.1 Motivation

In an ideal ecological sampling experiment one would proceed as follows. First the ecologist makes a list of species of interest, as well as a list of sites to be studied. For simplicity, assume that all sites satisfy the same environmental conditions. Then the ecologist takes one sample from each of the sites and counts the number of abundances of each species that appears on the list at each sampled site; it is assumed here that no errors in the counting process occur. If the individual organisms are also independently located, the counts for each species are expected to be distributed as a Poisson distribution in this ideal setting. Such zeroes appear when the environmental conditions make life possible, but still the species is not present at that site. The most extreme situation arises when a species cannot live under the given environmental conditions, then the zero count happens with probability one, which can be described by a degenerate Poisson distribution with all probability mass concentrated at zero. When plants or animals are not independently distributed over the habitat, for exam-

ple, as individuals occur in clumps, or when the counting process is error prone, the observed counts may show overdispersion as compared to what is expected with a Poisson model. In this case the negative binomial distribution may be a better choice; see, for example, **?** and **?** for a discussion on models for count data in ecology. In extreme situations the counting method may suffer from small sensitivity for the detection of rare species. This happens, for example, in metagenomics projects in which the abundances are calculated from massive parallel DNA sequencing [**?**]. The lack of sensitivity typically results in zero-inflation, i.e. more observed zeroes than expected under Poisson or negative binomial distributions. **?** give arguments to motivate the use of the negative binomial distribution for the analysis of count data from DNA sequencing, but the excess of zeroes is not discussed by them. Earlier we assumed that the ecologist only counts abundances of species that were listed prior to the start of the study. In practice, however, the ecologist will report every species that is observed at at least one of the sites. It often happens that a rare species is observed at only a minority of the sites. The $n \times s$ abundance matrix $\boldsymbol{Y}$ will then contain many zeroes for these rare species. This type of problem particularly occurs in microbial ecology studies in which a very large number of potential microorganisms are observed. Finally, we should be aware that the analyst can only model the mean abundence as a function of observed environmental data. It may thus well be that the absence of a species at a site cannot be explained by the observed conditions, but that it is caused by the unobserved unfaverable environmental status of the sampling site.

In conclusion, zero abundances may be generated by eight processes, which are summarised in Table 2.1. The *true zero* and *false zero* terminology is taken from **?**, but the specification of the first up to the fourth kind is added by us. The true and false

zeroes of first and second kind arise when the observed environmental variables contain all important information. Among these, the true zeroes may be well described by Poisson or negative binomial distributions (true I: degenerate distribution with all probability mass concentrated at the zero count), but the false zeroes may require ZADs. When the observed environmental variables miss important aspects of the environmental conditions (zeroes of the third and fourth kind), the count distributions may show overdispersion and, to some extent, also excess zeroes. True zeroes of the third kind (true III) form an example of the latter situation: despite the favorable observed environmental status, no species are observed because in reality the environmental conditions are bad. True zeroes of second and fourth kind happen when a species is absent at sites with favorable environmental conditions. Although this is in contradiction to the Baas-Becking hypothesis that states that "*everything is everywhere, but the environment selects*", studies have provided evidence that sometimes certain biogeographic patterns may cause a species to be absent at sites with good environmental conditions [**?**]. Note also that some of the causes of zero inflation (e.g. observer error) may also result in overdispersion, or even in underdispersion. Thus, when true and false zeroes are expected, a ZAD should be used. **?** and **?**, among others, discussed such models for analysing count data of rare species, but the effect of zero-inflation on CCA related methods has not been discussed yet in detail.

2.3.2 Zero-Inflated Models

A first class of ZADs is based on zero-inflated distributions. A first example is the zero-inflated Poisson (ZIP) distribution, which is a mixture of a Poisson distribution and the point probability at the zero count. The Poisson distributional assumption (2.1) is now

Table 2.1: Types of zero abundances in terms of the observed en true environmental conditions (F: favorable; U: unfavorable), the true absence or presence of the species, the species being selected and/or observed in the sample.

Type	species	env. cond. observed	env. cond. true	selected in sample	observed in sample
true I	absent	U	U		
true II	absent	F	F		
true III	absent	F	U		
true IV	absent	U	F		
false I	present	F	F	no	
false II	present	F	F	yes	no
false III	present	U	F	no	
false IV	present	U	F	yes	no

replaced with

$$Y_{ik}|\boldsymbol{x}_i \sim Y_{ik}|z_i \sim \text{ZIP}(\pi_{ik}, \lambda_{ik}),$$

($i = 1, \ldots, n; k = 1, \ldots, s$) where the parameter π_{ik} is the mixing parameter which is the probability of excess zeroes. See Appendix A.1.1 for more details on the ZIP distribution.

Because the zeroes may now also arise because of false zeroes, these zeroes may not contribute to the response function. When we assume that such zeroes occur completely at random, we may still consider $\lambda_{ik} = f_k(z_i)$, i.e. the response function only refers to the mean of the Poisson distribution in the ZIP mixture.

The ZIP model is appropriate when we may assume that the observed counts are measured without error. However, particularly with modern high throughput devices it may be expected that the abundances are obtained with measurement error [**??**]. As a consequence counts show more variability than the true abundances in the sample, and thus the Poisson behavior is corrupted in the sense that more variability is present in the counts than what is expected under the Poisson model. Other reasons are sum-

marised in, for example, **?**. The same arguments hold for the Poisson component in the ZIP distribution. The negative binomial (NB) distribution may be a more appropriate model, resulting in a zero-inflated negative binomial (ZINB) model,

$$Y_{ik}|\boldsymbol{x}_i \sim Y_{ik}|z_i \sim \mathsf{ZINB}(\pi_{ik}, \lambda_{ik}, \rho_{ik}),$$

($i = 1, \ldots, n$; $k = 1, \ldots, s$) where the parameter π_{ik} is the mixing parameter, $\lambda_{ik} = \mathrm{E}(Y_{ik}|z_i)$ is the mean of the NB distribution and ρ_{ik} is the overdispersion parameter of the NB distribution. As with the ZIP model, it is still appropriate to consider $\lambda_{ik} = f_k(z_i)$. See Appendix A.1.2 for more details on the ZINB distribution.

2.3.3 Hurdle Models

Another zero-altered Poisson distribution is the hurdle-zero-truncated-Poisson (HZTP) distribution. It may be considered as a computationally simpler alternative to the ZIP. The HZTP distribution may be defined as a mixture distribution of a point probability at the zero count and a zero-truncated Poisson (ZTP) distribution for the non-zero counts. The distributional assumption may be written as follows. For $i = 1, \ldots, n$; $k = 1, \ldots, s$,

$$Y_{ik}|\boldsymbol{x}_i \sim Y_{ij}|z_i \sim \mathsf{HZTP}(\pi_{ik}, \lambda_{ik}),$$

where π_{ik} is the probability of a zero count, and $\lambda_{ik} = \mathrm{E}(Y^*_{ik}|z_i)$, in which Y^*_{ik} is distributed as the (not-truncated) Poisson distribution. Appendix A.1.3 gives more details. The parameter π_{ik} is conventionally modelled by the environmental variables of interest through a logistic regression. However, in metagenomics, the probability of a zero

count for an OTU is known to be dependent on the library size, thus it is recommended to include the library size as covariate or offset to the linear combination of a set of environmental covariates.

The hurdle distribution is appropriate when zero counts may arise because of a true or a false zero. We particularly consider the HZTP model because of its computational advantages caused by the estimation orthogonality of the π and λ parameters. The response function is still meaningfully defined as $f_k(z_i) = \lambda_{ik}$.

Similarly as with the zero-inflated models the ZTP component may be replaced with a zero-truncated negative binomial (ZTNB) component when overdispersion is expected. See Appendix A.1.4 for more details

2.4 Estimation of the Environmental Gradient

The estimation of $\boldsymbol{\alpha}$ is complicated by the fact that all s models share the same $\boldsymbol{\alpha}$ vector. **?** proposed a maximum likelihood solution by extending the reduced-rank vectorised GLMs of **?** to cope with quadratic effects of z_i. This resulted in QRR-VGLM. Although our methods could also be developed in the QRR-VGLM framework, we prefer the computationally less demanding likelihood-ratio based methodology of **?**.

? proposed an iterative algorithm: (1) provide an initial $\boldsymbol{\alpha}$; (2) fit all s Poisson models separately; (3) reestimate $\boldsymbol{\alpha}$ by maximizing a likelihood ratio (LR) criterion (see further) using the maximum likelihood estimates from (2); (4) repeat steps (2) and (3) until

convergence. In the present context the likelihood-ratio criterion is defined as

$$\mathrm{LR}(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\sum_{k=1}^{s} y_{ik} \log\left(\frac{f_k(\boldsymbol{\alpha}^T\boldsymbol{x}_i)}{\sum_{j=1}^{s} f_j(\boldsymbol{\alpha}^T\boldsymbol{x}_i)}\right) - \sum_{i=1}^{n}\sum_{k=1}^{s} y_{ik}\log b_k, \tag{2.3}$$

where b_k is the relative frequency of species k. Because the last term in (2.3) does not depend on $\boldsymbol{\alpha}$, b_k has not to be specified or estimated. The theoretical development of the LR-based method is briefly described in Section 1.2.3. We also mention that it gives exactly the maximum likelihood solution of $\boldsymbol{\alpha}$ when a conditional multinomial model is considered[**?**]. The latter is obtained by starting from a Poisson distribution and conditioning on the total abundances at all sampling locations. **?** argued that this generates a competition among the species. The LR criterion keeps an attractive interpretation, even in the absence of the conditional multinomial distributional assumption: by the relationship between the LR criterion and the criterion used in Fisher's linear discriminant analysis (LDA), the $\boldsymbol{\alpha}$ found by maximising LR gives the maximal separation of the species response functions. The equivalence between CCA and LDA has been shown before [**??**]. In particular, under the CCA-assumptions of equal tolerances and Gaussian response functions, the estimate of $\boldsymbol{\alpha}$ is the direction in the p-dimensional environmental variables space along which the species vary most in terms of their optima μ_k ($k = 1, \ldots, s$). This characteristic was utilised by **?** as an alternative way for constructing the CCA solution. **?** give the LR criterion an LDA interpretation for a larger class of models. Finally, **?** conclude that their approach is flexible in the sense that the response functions may even be estimated nonparametrically.

When more than one environmental gradient is of interest, one typically proceeds as follows. First, the first environmental gradient, $\boldsymbol{\alpha}_1$, say, is estimated as described in the

previous paragraph. This gives n scores on the environmental gradient, $z_{1i} = \boldsymbol{\alpha}_1^T \boldsymbol{x}_i$ ($i = 1, \ldots, n$). Subsequently the p environmental variables are regressed on the first score z_1, i.e. p regression models are fitted,

$$x_{ij} = \beta_{0j} + \beta_{1j} z_{1i} + \varepsilon_{ij},$$

$j = 1, \ldots, p$, with ε_{ij} the mean zero error term and β_{0j} and β_{1j} the parameters. As the residuals, $e_{ij} = x_{ij} - \hat{\beta}_{0j} + \hat{\beta}_{1j} z_{1i}$, are orthogonal to the scores, they may be interpreted as transformed environmental variables that share no information with the first score and therefore they may replace the x_{ij} to find the second environmental gradient. In particular, the scores on the second environmental gradient are given by $z_{2i} = \boldsymbol{\alpha}_2^T \boldsymbol{e}_i$, where $\boldsymbol{\alpha}_2$ is again found by iteratively maximizing (2.3) and estimating the parameters in one of the models for the count data. More environmental gradients may be found by repeating this process. In particular, for finding the mth gradient residuals are obtained by fitting linear regression models with scores from all previous $m - 1$ gradients as regressors.

The results of a constrained ordination analysis can be graphically presented by an ordination plot. The ordination diagram is constructed as follows. The coefficients (loadings) of each environmental variable on the first two environmental gradients ($\boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_2}$) are displayed by arrows. Sites and species are plotted as points. The sites are naturally represented by their environmental scores and the species by their optima. Figure 2.1 shows an ordination diagram for a simulated data set (see Section 2.5.1 for details of the simulation settings). In particular, the plot shows that species 1-20 are well separated from species 21-40 along the first environmental gradient. It is also clear

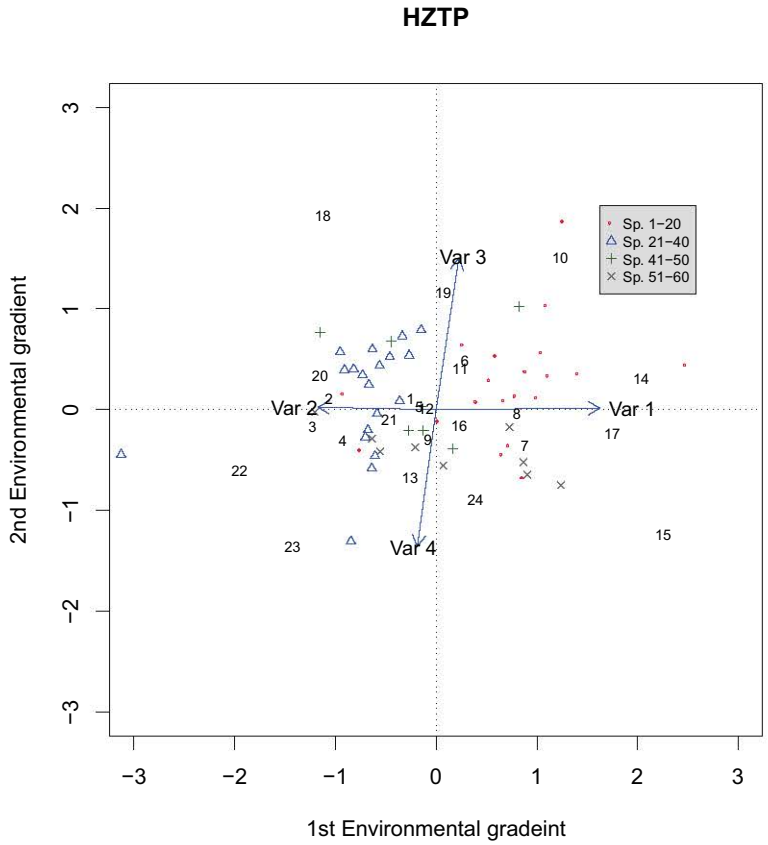


Figure 2.1: Ordination diagram for a simulated data set. Numbers represent 24 sampling locations. Symbols in different colors indicate the corresponding species optima.

that environmental variables 1 and 2 play leading roles in forming the first ordination scores. Variables 3 and 4 dominate the second ordination score.

2.5 Simulation Study

In this section we present results from three simulation studies. In the first experiment we examined the capability of the HZTP model to find the structure in data with zero inflation. The objective of the second simulation study is similar, but now a more complicated data structure is generated and several zero-altered distributions are evaluated and their performance is compared with the Poisson CGO and the ordinary CCA method. This study also includes overdispersed data. Next we empirically investigated

the effect of the frequency of zero abundances on the final results of a constrained ordination analysis. All simulations have been performed using the R software [**?**].

Note that the objective of the simulation studies is to empirically investigate the effect of zero-inflation on the behavior of the constrained ordination methods. It is not our intention to study all statistical properties of the estimators involved in the models. In this sense, the simulation study may seem incomplete.

2.5.1 Detecting a simple structure in the presence of zeroes

The objective is to assess to what extend the HZTP model succeeds in recognizing the structure in a data set that contains many zero abundances, and this is compared to other methods.

Data are simulated as follows.

1. **Environmental variables**: a subset of the environmental measurements collected by **?** is employed here to form a 24×4 environmental matrix $\boldsymbol{X}$. The first two variables have small correlations with the last two variables.

2. **Environmental gradients**: the two environmental gradients are set to orthogonal contrasts of the four environmental variables: $\boldsymbol{\alpha}_1^T = (1, -1, 0, 0)/\sqrt{2}$ and $\boldsymbol{\alpha}_2^T = (0, 0, 1, -1)/\sqrt{2}$. The scores are calculated as $\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{\alpha}_1$ and $\boldsymbol{z}_2 = \boldsymbol{X}\boldsymbol{\alpha}_2$.

3. **Non-zero abundances**: non-zero abundances for 60 species are generated using the random observations from a ZTP with $\lambda_{ik} = 30 - (z_{1i} - 2)^2$ for $k = 1, \ldots, 20$ (i.e. for the first 20 species), $\lambda_{ik} = 30 - (z_{1i} + 2)^2$ for $k = 21, \ldots, 40$ (i.e. for the second series of 20 species), $\lambda_{ik} = 30 - (z_{2i} - 2)^2$ for $k = 41, \ldots, 50$ (i.e. for the

third series of 10 species) and $\lambda_{ik} = 30 - (z_{2i} + 2)^2$ for $k = 51, \ldots, 60$ (i.e. for the last 10 species). This results in a 24×60 matrix, say $\boldsymbol{Y}$, with positive abundances.

4. **Zero abundances**: each element of the $\boldsymbol{Y}$ matrix is set to zero with a probability of $70\ \%$.

Using this data generation procedure 1000 data sets have been simulated. For each data set the environmental gradients have been estimated using CCA, the Poisson and HZTP based methods. The results are presented in Table 2.2. Note that the means have been normalised (i.e. $\boldsymbol{\alpha}^T\boldsymbol{\alpha} = 1$) so as to make the estimated vectors comparable. We conclude that the HZTP model results in much less biased estimates of the environmental gradients, particularly for the first gradient. Moreover, the standard deviations of the HZTP-based estimators are much smaller than those of the other estimators. CCA gives rather good estimates too, but only for the first dimension. This may in part be explained by the common tolerances that were used in the species response function for simulating the data.

The simulation results are graphically summarised in Figure 2.2. The graphs illustrate that the HZTP method succeeds well in terms of finding the underlying structure of the data as well as the relative importance of each environmental variable along either gradient. In particular, the plot of the HZTP model shows that species 1-20 are well separated from species 21-40 along the first environmental gradient. It is also clear thatc. Variables 3 and 4 dominate the second ordination score. These structures are less clear from the ordination diagrams for the Poisson model (Figure 2.2, right panel) and CCA (Figure 2.2, left panel).

	HZTP		Poisson		CCA	
Variables	m($\boldsymbol{\alpha}_1$)	sd($\boldsymbol{\alpha}_1$)	m($\boldsymbol{\alpha}_1$)	sd($\boldsymbol{\alpha}_1$)	m($\boldsymbol{\alpha}_1$)	sd($\boldsymbol{\alpha}_1$)
x_1	0.7470	0.0417	0.3225	0.5036	0.4665	-0.4671
x_2	-0.6513	0.0465	-0.5228	0.5261	-0.2633	0.6571
x_3	-0.0645	0.0527	0.0732	0.1860	0.0236	0.1797
x_4	-0.0300	0.0785	0.0212	0.2291	0.1132	-0.1347
	HZTP		Poisson		CCA	
Variables	m($\boldsymbol{\alpha}_2$)	sd($\boldsymbol{\alpha}_2$)	m($\boldsymbol{\alpha}_2$)	sd($\boldsymbol{\alpha}_2$)	m($\boldsymbol{\alpha}_2$)	sd($\boldsymbol{\alpha}_2$)
x_1	0.1357	0.1960	0.5092	0.5054	-0.4671	0.4888
x_2	0.1775	0.2119	-0.2073	0.5421	-0.4315	0.5158
x_3	0.4652	0.1259	0.1023	0.2293	0.0656	0.2782
x_4	-0.7795	0.1640	0.0739	0.2839	-0.1009	0.2088

Table 2.2: The normalised means (m) and standard deviations (sd) of the elements of the $\boldsymbol{\alpha}$ vector estimates obtained with the Poisson, HZTP and CCA methods. The true $\boldsymbol{\alpha}$ vectors were $(0.707, -0.707, 0, 0)^T$ and $(0, 0, 0.707, -0.707)^T$.

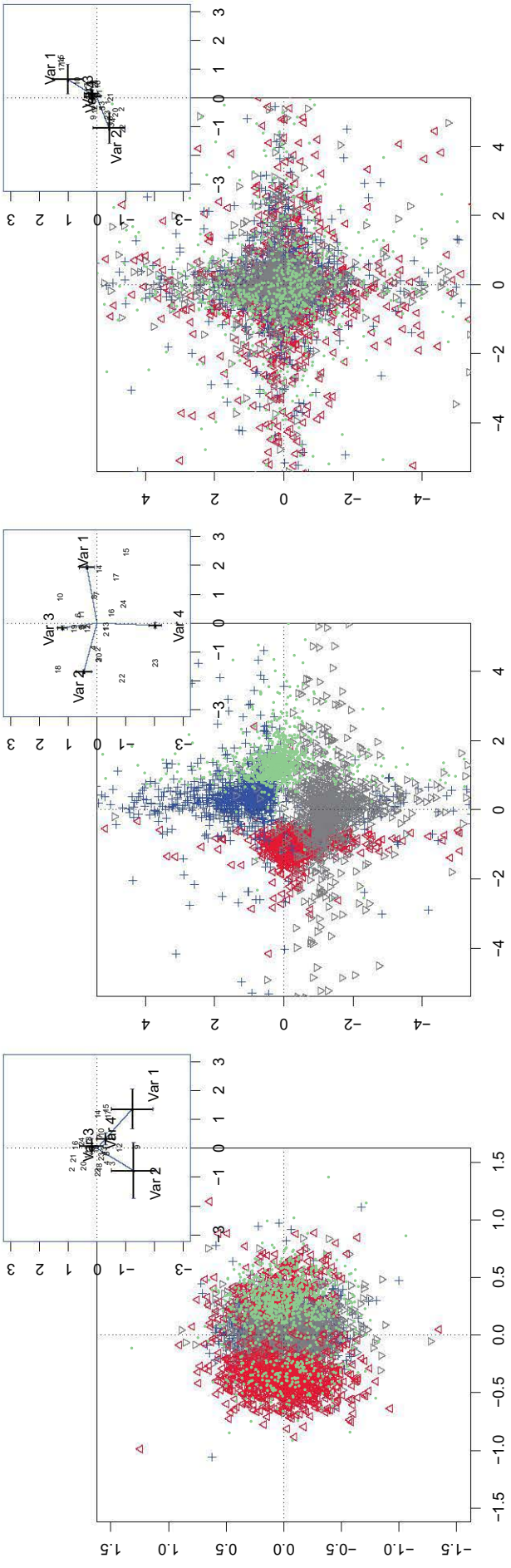


Figure 2.2: Ordination diagrams of the CCA (left panels), HZTP (middle panels) and the Poisson (right panels) solutions. The lower panels show the distribution of the optima of four species (one species of each of the four species groups). The top panels show the averages of the gradients over the 1000 simulation runs. The horizontal and vertical bars indicate the standard deviations. The sites are represented by the linear combination of simulated environmental variables and average gradient on both dimensions.

2.5.2 Detecting a complicated structure in the presence of zeroes

The set up of this simulation study is similar to the previous study, except that now a more complicated, but more realistic structure is used for the simulation of the abundances. In particular, the environmental data ($\boldsymbol{X}$) is a 44×13 real data matrix from a project in which we are involved [**?**]. The two gradients are set to the estimates obtained from that same study. Non-zero abundances for 100 species are simulated using the ZTP distribution with Gaussian response functions as in (2.2) with parameters set to the estimates that we found in the project. The π parameter of the HZTP distribution is related to the environmental variables in a quadratic logistic regression model of which the parameters are again set to the estimates found in the project. This procedure resulted in about 50% zero abundances on average. For each generated data set the gradients were estimated with CCA and the Poisson, HZTP and ZIP based methods. The same simulation procedure is repeated with the ZTP data generating distribution replaced with the ZTNB distribution.

The results of the first gradient estimation are presented in Tables 2.3 and 2.4. The HZTP model succeeds well in estimating the true relation, whereas the other methods show substantial bias. Also in terms of the standard deviations of the estimators, the HZTP model gives the best results. Note that CCA now gives worse results than the other methods. This may be explained by the nonconstancy of the tolerances for the given data set.

		ZTP				ZTNB			
Variables	α	CCA	Poisson	HZTP	ZIP	CCA	Poisson	HZTP	ZIP
Depth	0.0049	**-0.0069**	**0.0155**	0.0048	**0.0122**	**-0.0098**	**0.0132**	0.0048	**0.0105**
pH	0.4154	**-0.2406**	**0.3558**	0.4140	**0.3507**	**-0.2436**	**0.3563**	0.4140	**0.3496**
Conductivity	0.2470	**-0.2095**	0.2666	0.2472	0.2566	**-0.2217**	0.2631	0.2466	0.2484
TOC	-0.0019	**-0.0045**	**0.0345**	**-0.0009**	**0.0261**	**-0.0220**	**0.0322**	-0.0019	**0.0247**
DOC	0.0613	**-0.0267**	**0.0116**	0.0601	**0.0151**	**-0.0164**	**0.0138**	0.0607	**0.0129**
Na	0.1204	**-0.0631**	**0.0862**	0.1200	**0.0904**	**-0.0510**	**0.0882**	0.1203	**0.0941**
K	0.4974	**-0.3075**	0.4650	0.4979	0.4627	**-0.3044**	0.4620	0.4950	0.4603
Ca	0.1114	**0.0814**	0.1091	0.1117	0.1067	**-0.0847**	0.1089	0.1109	0.1079
Mg	-0.1720	**0.0822**	**-0.1176**	-0.1708	**-0.1238**	**0.0690**	**-0.1205**	-0.1729	**-0.1247**
Cl	-0.1406	**0.0968**	**-0.1132**	-0.1398	**-0.1116**	**0.0965**	**-0.1156**	-0.1404	**-0.1128**
SO4	0.5191	**0.3205**	0.4738	0.5186	0.483 0	**-0.3230**	0.4715	0.5175	0.4804
NH4-N	0.1746	**-0.1040**	**0.1518**	0.1739	**0.1410**	**-0.0972**	**0.1516**	0.1738	**0.1416**
Silicate-Si	-0.3729	**0.2780**	-0.3859	-0.3725	-0.3644	**0.2953**	-0.3839	-0.3725	-0.3584

Table 2.3: The means of the simulated estimates of the α_1 vectors obtained from CCA, and the Poisson, HZTP and ZIP based methods when the data are simulated with the ZTP and the ZTNB models. The elements in bold indicate the estimates that show more than 10% bias. The results are based on 1000 simulations.

	ZTP				ZTNB			
Variables	CCA	Poisson	HZTP	ZIP	CCA	Poisson	HZTP	ZIP
Depth	0.1267	0.0812	0.0099	0.0874	0.1290	0.0815	0.0184	0.0928
pH	0.1892	0.0905	0.0108	0.1032	0.1818	0.0942	0.0196	0.1108
Conductivity	0.1429	0.0901	0.0120	0.1076	0.1528	0.0920	0.0199	0.1065
TOC	0.2715	0.1710	0.0237	0.1636	0.2743	0.1749	0.0342	0.1747
DOC	0.2257	0.1909	0.0260	0.1833	0.2278	0.1935	0.0370	0.1980
Na	0.2410	0.0744	0.0089	0.0826	0.2410	0.0757	0.0169	0.0799
K	0.2487	0.0987	0.0122	0.1203	0.2413	0.1032	0.0191	0.1233
Ca	0.1111	0.0782	0.0094	0.0866	0.1108	0.0824	0.0162	0.0930
Mg	0.2952	0.0957	0.0117	0.0995	0.2906	0.0963	0.0192	0.1061
Cl	0.1687	0.0850	0.0108	0.0903	0.1627	0.0887	0.0170	0.0966
SO4	0.2509	0.0690	0.0075	0.0868	0.2404	0.0714	0.0131	0.0908
NH4-N	0.1997	0.0657	0.0089	0.0808	0.1970	0.0693	0.0165	0.0791
Silicate-Si	0.1957	0.0823	0.0105	0.0921	0.1938	0.0827	0.0163	0.0902

Table 2.4: The standard deviations of the simulated estimates of the α_1 vectors obtained from the CCA, Poisson, HZTP and ZIP models when the data are simulated with the ZTP and the ZTNB models. The results are based on 1000 simulations.

2.5.3 The effect of the frequency of zero abundances

To assess the effect of the frequency of zero abundances in a realistic setting, we performed an in-silico experiment in which we started from a real data set with not too many zeroes, and we gradually added zeroes while observing the behavior of the gradient estimation methods. We considered the hunting spider data that was also analysed by **?**, among others. The data set contains six environmental variables for 28 sites in the Dutch dunes. Abundances of 12 hunting spider species have been recorded. The spider species *Trochosa terricola* has been selected for this empirical study. In the original data set this species had the most nonzero abundances. In 16 steps these abundances are replaced by zeroes. In particular, in the first step the smallest nonzero abundance of *Trochosa terricola* is replaced with zero, and the analyses are performed using CCA, the Poisson and the HZTP models. In the next step the next smallest nonzero abundance is replaced, and the data are re-analysed. This is repeated until four nonzero *Trochosa terricola* abundances remain. To measure the effect of adding zeroes, we computed a distance measure in each of the 16 steps. This is explained in the next paragraph.

Referring to (2.2), we use $\hat{\mu}_{kdm}$ to denote the estimated optimum of species k in the direction of the dth environmental gradient ($d = 1, 2$) in step $m = 0, \ldots, 16$ of the zero adding procedure, with $\hat{\mu}_{kd0}$ the estimated optimum of the original data set. For each species k and each step m the distance from the original optimum estimate in the dimension of the dth environmental scores is calculated as

$$d^2_{kdm} = (\hat{\mu}_{kdm} - \hat{\mu}_{kd0})^2.$$

The total deviation from the original species optimum is then calculated as

$$d_m = \sum_{k=1}^{12} \sqrt{(d_{k1m}^2 + d_{k2m}^2)}.$$

These deviations may also be computed in each direction separately, i.e. $d_{dm} = \sum_{k=1}^{12} \sqrt{d_{kdm}^2}$. Figure 2.3 shows the results. This graph clearly demonstrates that an increasing frequency of zeroes has a greater effect on the Poisson solution than on the solutions of CCA and HZTP.

Figure 2.4 shows the ordination diagrams of all three methods on the original data, whereas Figure 2.5 shows the diagrams of them in step 16, after having added 22 zero abundances to *Trochosa terricola* species. There are many differences that can be observed by comparing these plots; we only discuss a few. On the original data, Figure 2.4 shows that the Poisson and HZTP methods give comparable results, but the CCA method gives different results, even after reflection of the axes. After having introduced 22 zeroes Figure 2.5 shows even quite some more differences between the methods, particularly the Poisson solution changed noticeable. For example, in the middle panel of Figure 2.4, species *Aulonia albimana* seems to be most abundant at sites 6, 7 and 13, yet the diagram in the right panel suggests that this species is most abundant at sites 2, 4 and 7. When consulting the original complete data set, we see that the spider *Aulonia albimana* has the largest abundances (30 and 24) at sites 2 and 4. Thus, the HZTP-based method provides a more correct conclusion w.r.t. *Aulonia albimana*. Another example: after the introduction of 22 zeroes to the species *Trochosa terricola*, the diagram of the HZTP model still shows the correct conclusion, but not the plots of the Poisson and CCA methods. Many more such examples can

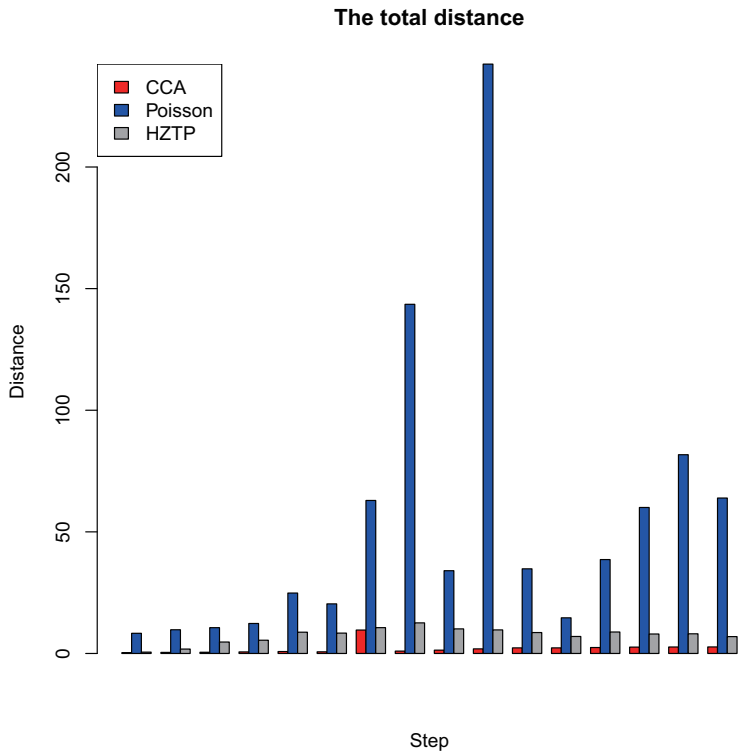


Figure 2.3: The total distance d_m in each step $m = 0, 1, \ldots, 16$.

be read from the figures. From this simulation experiment we again conclude that the HZTP model is more robust to an increasing frequency of zero abundances.

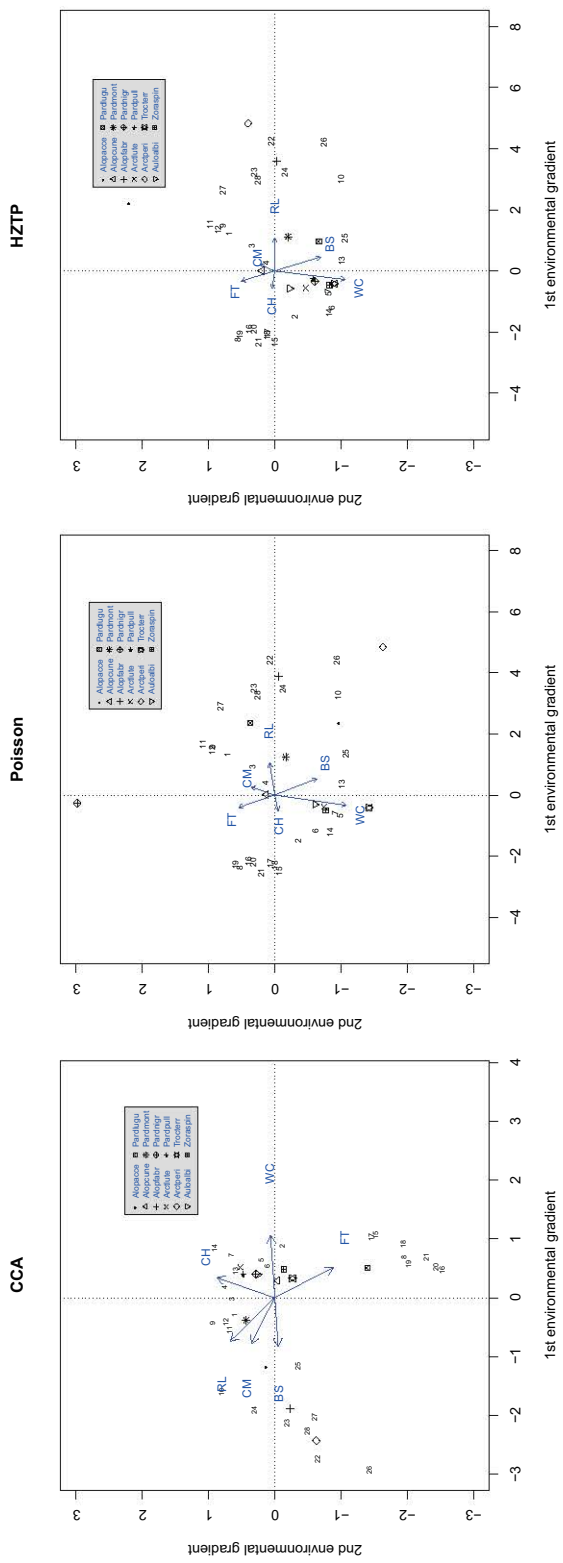


Figure 2.4: Ordination diagrams for the hunting spider data, obtained with CCA (left), the Poisson (middle) and the HZTP (right) models. Numbers indicate sites and the abbreviation of species name in the legend are: Alopacce (*Alopecosa accentuata*), Alopfabr (*Alopecosa fabrilis*), Arctlute (*Arctosa lutetiana*), Arctperi (*Arctosa perita*), Auloalbi (*Aulonia albimana*), Pardlugu (*Pardosa lugubris*), Pardmont (*Pardosa monticola*), Pardnigr (*Pardosa nigriceps*), Pardpull (*Pardosa pullata*), Trocterr (*Trochosa terricola*) and Zoraspin (*Zora spinimana*). The environmental variables are *BS* (percent cover of bare sand), *CM* (percent cover of the moss layer), *CH* (percent cover of the herb layer), *FT* (percent cover of fallen leaves and twigs), *RL* (reflection of the soil surface with cloudless sky) and *WC* (percentage of soil dry mass).

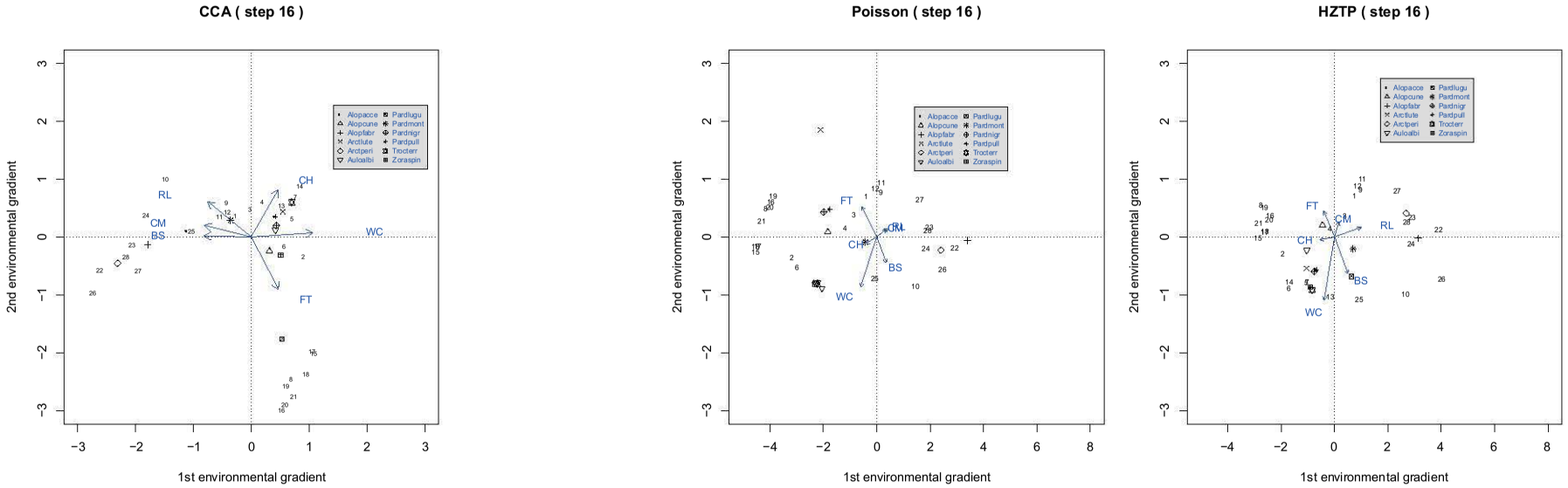


Figure 2.5: Ordination diagrams for the hunting spider data, obtained with CCA (left), the Poisson (middle) and the HZTP (right) models, after introducing 22 zero abundances sequentially. The abbreviations of species and environmental variables are as in Figure 2.4.

2.6 Absence/Presence Modelling

2.6.1 A binomial model

When looking at the LR criterion (2.3) in detail for the HZTP distribution with response function $f_k(z_i) = \lambda_{ik}$, it becomes clear that only the ZTP component is involved and that only the non-zero abundances are required in the computations. The hurdle part with parameter π_{ik}, is not used at all. The parameter π_{ik}, however, can also be linked to the environmental scores using the binomial distribution and, for example, the logit link. In particular,

$$\pi_{ik} = \mathrm{P}\left(Y_{ik} > 0 | z_i\right) = \text{expit}\left(a_k - \frac{(z_i - \mu_k)^2}{2t_k^2}\right), \tag{2.4}$$

where the parameters in the quadratic model are as in (2.2). The parameter μ_k has still the interpretation of the optimum of species k, i.e. it is the environmental score of sites for which species k has the largest probability of being present. For the estimation of the parameters in this logistic regression model, it is sufficient to work with the absence/presence data. This model resembles the *Gaussian logit model* of **?** and the QRR-VGLM model of **?**, in which the logit-transform of probability is also modeled as a quadratic function of environmental variables. Abundance data sets may be very unbalanced with respect to the absence/presence data: some species are (almost) always present, whereas others are (almost) nowhere present. This might result in the *separation* problem. We have used the method of **?** to resolve the problem.

Instead of estimating the environmental gradient $\boldsymbol{\alpha}$ by maximizing (2.3) based on the ZTP response function, one may choose to estimate $\boldsymbol{\alpha}$ by maximizing a LR criterion

based on the binomial model. In its most general form, the log likelihood ratio criterion expresses the log ratio of two maximized likelihoods [**?**]. The first is the likelihood of the data given separate models for each species, whereas the second is the likelihood of the data given a common model for all species. On using the binomial models, and writing $\pi_k(\boldsymbol{\alpha}^T\boldsymbol{x}_i)$ for (2.4), we find

$$\text{LR}(\boldsymbol{\alpha}) = \log \prod_{i=1}^{n}\prod_{k=1}^{s} \pi_k(\boldsymbol{\alpha}^T\boldsymbol{x}_i)^{y_{ik}}(1-\pi_k(\boldsymbol{\alpha}^T\boldsymbol{x}_i))^{1-y_{ik}} - \log \prod_{i=1}^{n}\prod_{k=1}^{s} \pi(\boldsymbol{\alpha}^T\boldsymbol{x}_i)^{y_{ik}}(1-\pi(\boldsymbol{\alpha}^T\boldsymbol{x}_i))^{1-y_{ik}},$$

where $\pi(z) = \text{expit}\left(a - \frac{(z-\mu)^2}{2t^2}\right)$, which represents the common model.

2.6.2 Data example

The hunting spider data is again used to illustrate the method. For this purpose the spider data has been converted into 0-1 indicators. The common model parameter $\pi(z)$ is specified as the weighted average of the probabilities estimated from the separate models, i.e. $\pi(z_i) = \sum_{c=1}^{s} \frac{y_{+c}}{y_{++}}\pi_c(z_i)$).

Again an ordination diagram type of graph may be plotted and used for summarising the results of the analysis. For example, Figure 2.6 shows that site 12, together with its surrounding sites (1, 9, and 11), is probably more preferable for species *Alopecosa accentuata*, as compared to sites 16-20. Note that the signs of the first environmental gradient agrees with the results from the HZTP analysis (see Figure 2.4).

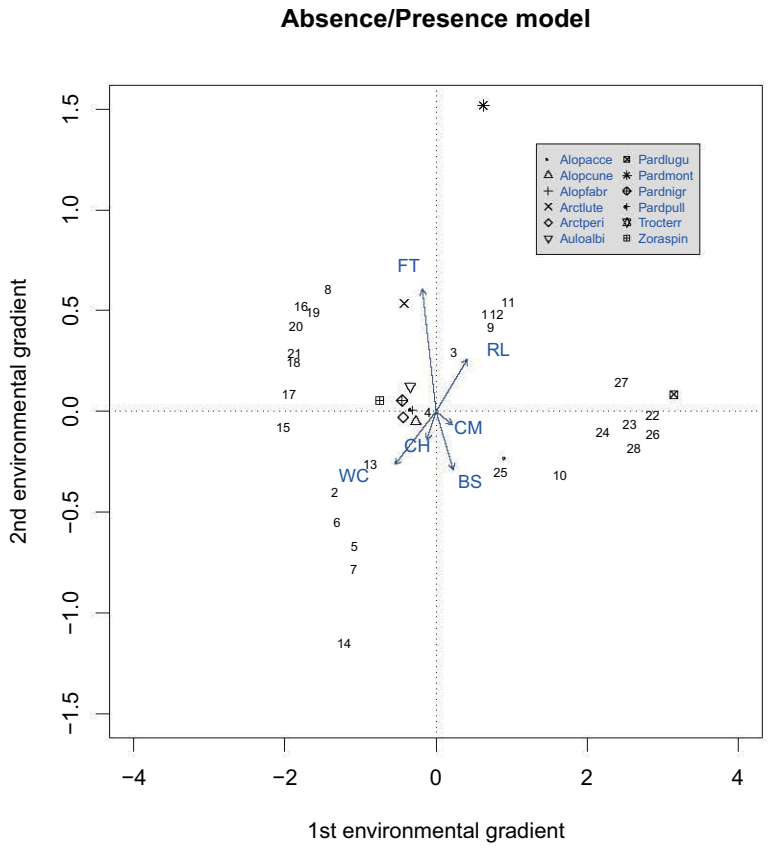


Figure 2.6: The ordination diagram for the analysis of the absence/presence hunting spider data. See the legend of Figure 1 for the abbreviations of species and environmental variables.

2.7 Discussion

We have presented a constrained ordination analysis method that deals properly with zero inflation in abundance studies. Our method consists basically in replacing the Poisson distribution with ZADs. In particular, we propose the ZIP, ZINB, HZTP and the HZTNB distributions. From several simulation experiments we have concluded that CCA and the Poisson based method give more biased results in the presence of zero-inflation, whereas the use of the ZADs still results in correct conclusions. Based on the empirical comparison of methods we conclude that the HZTP is preferred over the ZIP model. Moreover, the parameter estimation for the HZTP is computationally more convenient.

Both the zero-inflated models and the hurdle models have a parameter that refers to the presence of excess zeroes. This parameter may also be modeled through a logistic regression model with the environmental scores as predictors. This has been proposed before by **?** and **?**. We have illustrated this method for the π parameter in the hurdle models, but it can also be applied to the π parameter of the zero-inflated models. In the latter case the estimation of the parameters in the hurdle part and in the Poisson part could be performed simultaneously.

We conclude that ZADs should be used in the presence of too many zero abundances. The HZTP seems to work well, even with a little overdispersion in the data. We propose to estimate the environmental gradients for both the ZTP and the hurdle part of the HZTP model. This results in two informative ordination diagrams that may help the ecologist to gain insight into the data.

All computations were performed using the R software [**?**], furthermore, we used the

R function ‘vglm.fit’ from **?** for fitting Zero-truncated models and Zero-inflated models. Also the R package ‘snowfall’ [**?**] is employed for parallel computing so as to reduce the computational time. The R functions are available from the authors on simple request.

Chapter 3

Constrained Ordination Analysis with Enrichment of Bell-Shaped Response Functions

Summary: Constrained ordination methods aim at finding an environmental gradient along which the species abundances are maximally separated. The species response functions, which describe the expected abundance as a function of the environmental score, are according to the ecological fundamental niche theory only meaningful if they are bell-shaped. Many classical model-based ordination methods, however, use quadratic regression models without imposing the bell-shape and thus allowing for meaningless U-shaped response functions. The analysis output (e.g. a biplot) may therefore be potentially misleading and the conclusions are prone to errors. In this paper we present a log-likelihood ratio criterion with a penalisation term to enforce more bell-shaped response shapes. We report the results of a simulation study and apply

our method to metagenomics data from microbial ecology.

Key words: biplot; correspondence analysis; ecology; multivariate statistics; penalisation.

3.1 Introduction

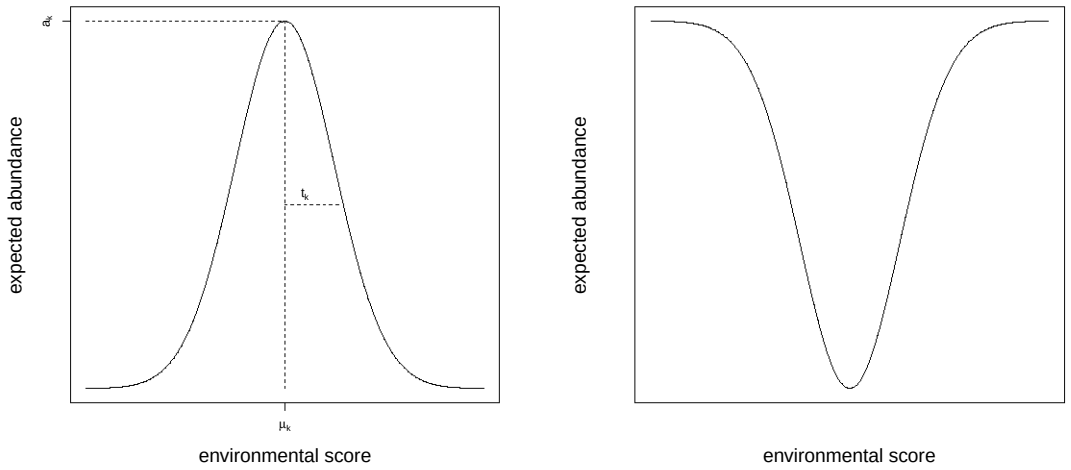
Constrained or Canonical Correspondence Analysis (CCA) is a well established method among environmental ecologists to study the relation between species abundances and the environmental conditions. This method, which is nowadays looked at as a particular technique for Constrained Ordination Analysis (COA), was originally introduced by **?** as an extension of the correspondence analysis method for exploring the dependence structure in a two-way contingency table or a species-by-sample abundance table; see, for example, **???** for recent accounts.

In a CCA and COA the species abundances are regressed on environmental scores that result from linear combinations of the environmental conditions of the sampling locations. The linear combination is referred to as the environmental score whose coefficients are referred to as environmental gradient. The coefficients defining the environmental score are the same for all species, but each species has its abundances described by another *response function* that relates the abundances to the environmental scores. The CCA method aims at finding the gradient that maximally separates the response functions. When the species-specific response functions are described by a Gaussian density function and the abundances are assumed to follow a Poisson distribution, a Constrained Gaussian Ordination (CGO) model arises. **?** showed that

his eigenvalue-based CCA can be considered as an approximation to the maximum likelihood solution of a CGO model. The approximation was studied in a simulation study by **?**; they concluded that the approximation is good when the species-specific tolerances are more or less equal (see later). While the eigenvalue-approach of CCA has the advantage of being computationally more efficient, the model-based likelihood method has the advantage of being more flexible as it allows for replacing the Poisson and Gaussian components with other functional forms, numerical instabilities notwithstanding.

The species response function is an important concept in community ecology and particularly for the quantitative analysis of ecological niches. Just like the CCA and CGO methods, the new method developed in this paper is appropriate for studying the *realised niches* or the *fundamental niches* of species. The latter can be defined as the range of environmental conditions under which the species can exist without interspecies competition or predation from other species [**?**]. Within a fundamental niche, the species response function has a single maximum that corresponds to the optimal environmental conditions for the existence of the species. When moving away from the optimal condition the abundances are expected to decrease; the relation is typically non-linear [**??**]. **?** developed a Gaussian ordination method to reveal the species-environment relation in which each species response function was assumed to be a unimodal bell-shaped curve, called a Gaussian curve; this is illustrated in the left panel of Figure 3.1. The response function is characterised by three parameters: the maximum a_k, optimum μ_k and the tolerance t_k. In 1986 **?** introduced the CCA method which soon became the dominating ordination analysis method in community ecology. The method is based on Gaussian response curves with equal tolerances.

Figure 3.1: a Gaussian response curve (left) and a U-shaped response curve (right). The left panel also indicates the parameters μ_k (optimum), a_k (maximum) and t_k (tolerance).



The *realised niche* of a species is a subset of the fundamental niche, and describes the environmental conditions under which the species can exists with interspecies competition and predation from other species. The symmetric unimodal response function has been criticised, particularly for studying the realised niches of a species. The competition among species may cause unimodal response functions of different skewness [**?**]. Many studies have focused on the use of more flexible species response curves. For example, **??** suggested that the β-function is a more realistic unimodal response function. However, computational difficulties in estimating the β-function parameters obstruct the application of β-functions in ordination analysis [**??**]. For the same reasons **?** proposed the HOF model which contains a set of submodels allowing different degrees of skewness and flatness. Note that the most complex HOF model has 5 parameters, causing the HOF model to suffer from similar computational obstacles as the β-function. Besides the technical difficulties, another reason that these more complex models are not very popular may be that bell-shaped curves are a realistic reflection of

the fundamental niche theory and in many ecological studies the unimodal bell-shaped response functions show sufficient approximation to the data.

Some papers [**??**] report attempts to embed ordination methods in classical statistical modelling frameworks, allowing for more flexibility in describing the response functions. For example, the work of **?** enables modelling each individual species in the community with maximal flexibility, both parametrically and nonparametrically. They proposed a likelihood ratio statistic that measures the separation of species response functions. Finding the gradient along which the species display maximally separated response functions is obtained by maximising their likelihood ratio criterion. Their method will be referred to as Flexible Constrained Ordination Analysis (FCOA).

Despite the flexibility of the FCOA method, users should be careful because any type of species response curve could be fitted, even if it is ecologically meaningless. For instance, one could use the popular second order polynomial Poisson regression model for relating the expected abundance to the environmental scores, but without any constraints on the polynomial regression coefficients the fitted curve can be either bell-shaped (concave) or U-shaped (convex). This is illustrated in Figure 3.1. Consequently, the result may be misleading, for the U-shaped response functions are often ecologically meaningless, particularly in the niche theory. This raises the question whether the contributions made by these U-shaped functions to the likelihood ratio criterion should not be counted at all, as they may eventually lead to the wrong or suboptimal environmental gradient. The problem becomes even worse when observing that many ordination software implementations do not allow the user to assess the quality and relevance of the model fit. Instead the ordination results are typically summarised in a biplot which shows the environmental gradients and the estimated optimum parameters

(μ_k) for all species. However, when a species μ_k parameter comes from a U-shaped response function, it does not represent the optimal environmental conditions, but quite the opposite. From the biplot, conclusions are formulated about similarities/differences between species responses to environmental conditions, assuming that all species show ecologically meaningful bell-shaped response functions along the environmental gradients.

In this paper we propose a penalized maximum likelihood method in which the penalization forces many response functions to be bell-shaped. We refer to the new method as BECOA (Bell-shape Enriched COA). The method is empirically evaluated in a simulation study and it is applied to the microbial diversity data. The paper ends with a discussion and the formulation of conclusions.

3.2 Bell-Shape Enriched Constrained Ordination Analysis

3.2.1 Model-Based Constrained Ordination Analysis

Our method builds on the construction of **?** for constrained ordination. First some notation is introduced. Let $\boldsymbol{x}_i^T = (x_{i1}, \ldots, x_{ip})$ denote the p-dimensional vector of environmental variables measured at sampling location $i = 1, \ldots, n$. The vector with the abundances of species $k = 1, \ldots, s$ at the n locations is denoted by $\boldsymbol{Y}_k^T = (Y_{1k}, \ldots, Y_{nk})$. For a given coefficient vector $\alpha^T = (\alpha_1, \ldots, \alpha_p)$, the environmental observations are transformed to univariate environmental scores given by $z_i = \boldsymbol{\alpha}^T \boldsymbol{x}_i$, $i = 1, \ldots, n$. The

abundance of species k at location i is assumed to be Poisson distributed, conditional on z_i. In particular,

$$Y_{ik}|\boldsymbol{x}_i \sim Y_{ik}|z_i \sim \text{Poisson}(\lambda_{ik}),$$

where $\lambda_{ik} = \mathrm{E}\left(Y_{ik} \mid \boldsymbol{x}_i; \boldsymbol{\beta}_k\right) = \mathrm{E}\left(Y_{ik} \mid z_i; \boldsymbol{\beta}_k\right)$ in which $\boldsymbol{\beta}_k$ is a regression parameter vector. The probability mass function is denoted by $p_k(y \mid z, \boldsymbol{\beta}_k)$. In the present context the Poisson mean function is referred to as the response function of species k, denoted by $f_k(z_i; \boldsymbol{\beta}_k)$, or simply $f_k(z_i)$.

If we assume that all species respond in the same way to the environment, the index k may be dropped from the response function, resulting in a common response function, say $f(z_i)$. The corresponding Poisson probability mass function is denoted by $p(y \mid z; \boldsymbol{\beta})$, and the $\boldsymbol{\beta}$ parameter may be estimated using data of all species simultaneously.

? proposed an iterative scheme for the joint estimation of $\boldsymbol{\alpha}$ and the $\boldsymbol{\beta}_k$ parameters: (1) for an initial $\boldsymbol{\alpha}$ the environmental variables are transformed to the z scores; (2) estimate the $\boldsymbol{\beta}_k$ and $\boldsymbol{\beta}$ from the corresponding Poisson regression models; (3) estimate $\boldsymbol{\alpha}$ by maximising the log-likelihood ratio (LLR) criterion

$$\text{LLR}(\boldsymbol{\alpha}) \;=\; \sum_{i=1}^{n}\sum_{k=1}^{s} \log \frac{p_k(y_{ik} \mid z_i; \boldsymbol{\beta}_k)}{p(y_{ik} \mid z_i; \boldsymbol{\beta})} \tag{3.1}$$

with the $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_k$ parameters replaced by their estimates from step (2); (4) repeat steps (2) and (3) until convergence. The rationale for the use of the LLR criterion is given later in this section.

The response function is often believed to be unimodal. A popular choice is the Gaussian response model, which borrows its name from its similarity to a Gaussian density

function. Upon using the canonical log-link function of Poisson regression, the model can be written as

$$\log f_k(z_i) = a_k - \frac{(z_i - \mu_k)^2}{2t_k^2} \tag{3.2}$$

in which $\boldsymbol{\beta}_k^T = (a_k, \mu_k, t_k)$ represents the species-specific parameters to be estimated from the data. Despite the convenient interpretation of the parameters, model formulation (3.2) is nonlinear in the parameters and hence computationally less attractive. Model (3.2) is therefore often replaced by the log-linear model

$$\log f_k(z_i) = \beta_{0k} + \beta_{1k} z_i + \beta_{2k} z_i^2 = \boldsymbol{\beta}_k^T \boldsymbol{w}_i, \tag{3.3}$$

with $\boldsymbol{\beta}_k^T = (\beta_{0k}, \beta_{1k}, \beta_{2k})$ and $\boldsymbol{w}_i^T = (1, z_i, z_i^2)$. For a given $\boldsymbol{\alpha}$, this model allows for the use of standard GLM software for the estimation of the β parameters. However, with $\hat{\boldsymbol{\beta}}_k^T = (\hat{\beta}_{0k}, \hat{\beta}_{1k}, \hat{\beta}_{2k})$ denoting the vector with the parameter estimates for species k, model (3.3) results in a U-shaped response function if $\hat{\beta}_{2k} > 0$.

The rationale of the use of the LLR criterion (3.1) can be understood as follows. A large LLR indicates that it is advantageous to model each species with a separate model (p_k). Hence, maximizing LLR(α) as a function of α will result in the environmental gradient that maximally separates the species-specific models. By formulating the criterion in terms of general distribution functions p_k, a very flexible method arises. Here we restrict the p_k's to be Poisson distributions, but e.g. **?** studied the use of the zero-inflated Poisson (ZIP) distribution and **?** even replaced p_k with nonparametric estimates. In this paper, however, the method is confined to the Poisson distribution and the Gaussian response function so that our method can be easily interpreted as an extension of the traditional ordination methods. The formulation of the constrained ordination method

in terms of likelihoods allows extending the method by making use of the rich theory of likelihood-based methods. In this paper, we suggest to estimate the β_{2k} parameters of model (3.3) with a penalised maximum likelihood method that favors negative parameter estimates and hence results in bell-shaped response functions. Details are given in the next section.

Although our methods are not restricted for use in the quadratic model (3.3), they will be demonstrated with model (3.3). In general we say that $\boldsymbol{\beta}_k$ and $\boldsymbol{w}_i$ are q-dimensional.

3.2.2 Penalised Maximum Likelihood

The penalised maximum likelihood estimation procedure is inspired by a Bayesian setup. Let $g(\boldsymbol{\beta}_k)$ denote the prior distribution on $\boldsymbol{\beta}_k$. Although we only require penalisation on β_{2k} we will present the method more generally so that penalisation on other β parameters becomes also possible. The penalised maximum likelihood estimator is then defined as $\boldsymbol{\beta}_k$ that maximises the posterior $p_k(\boldsymbol{\beta}_k \mid y, z) \propto p_k(y \mid z, \boldsymbol{\beta}_k) g(\boldsymbol{\beta}_k)$.

Penalised Poisson regression has been described before in the literature [**??**], but the focus is often on the L1 (lasso) and L2 (ridge) penalties that result from a Laplace and a normal prior, respectively. These priors have zero means which leads to estimates that are shrunken towards zero. In the present setting, however, we want to favour negative parameter values. Therefore we opt for a normal distribution with a negative mean.

We will describe two algorithms. The first algorithm relies on standard software for the estimation of the β parameters, whereas the second algorithm is an alternative formulation of the iterative reweighted least squares procedure that often forms the

core of the estimation algorithm in statistical software.

Algorithm 1

Consider the score equation for parameter β_{jk},

$$\sum_{i=1}^{n} \frac{\partial \log p_k(\boldsymbol{\beta}_k \mid y_{ik}, z_i)}{\partial \beta_{jk}} = \sum_{i=1}^{n} (y_{ik} - \exp(\boldsymbol{\beta}_k^T \boldsymbol{w}_i)) w_{ij} + n \frac{\partial \log g(\boldsymbol{\beta}_k)}{\partial \beta_{jk}} = 0. \tag{3.4}$$

We propose the following algorithm:

1. estimate $\boldsymbol{\beta}_k$ from the non-penalised regression. Denote this estimate by $\tilde{\boldsymbol{\beta}}_k$;
2. let $\tilde{\omega}_{ik} = -n \frac{\partial \log g(\tilde{\beta}_k)}{\partial \beta_{jk}} / w_{ij}$ and set $\tilde{y}_{ik} = y_{ik} - \tilde{\omega}_{ik}$;
3. find $\hat{\beta}_{jk}$ from solving the score equation $\sum_{i=1}^{n} (\tilde{y}_{ik} - \exp(\boldsymbol{\beta}_k^T \boldsymbol{w}_i)) w_{ij} = 0$;
4. set $\tilde{\boldsymbol{\beta}}_k = \hat{\boldsymbol{\beta}}_k$;
5. iterate through steps 2-4 until convergence.

When the prior g is a normal distribution with mean $\boldsymbol{\delta}^T = (\delta_0, \delta_1, \delta_2)$ and covariance matrix $\gamma^{-1}\boldsymbol{D}$ with γ a tuning parameter and $\boldsymbol{D}$ a diagonal matrix with elements d_0, d_1, d_2, we find $\tilde{\omega}_k = \gamma \sum_{j=0}^{2} (\beta_{jk} - \delta_j)/(w_{ij} d_j^2)$. Since we only require penalisation for β_{2k} we set $d_0 = d_1 = +\infty$ (in the limit), resulting in $\tilde{\omega}_k = \gamma(\beta_{2k} - \delta_2)/(w_{i2} d_2^2)$. In this case γ/d_2^2 may be replaced by a single penalisation parameter, e.g. by setting $d_2 = 1$. The prior parameter δ_2 must be negative to favour negative $\hat{\beta}_{2k}$'s.

Algorithm 2

Equation (3.4) may be written as (for all β_{jk}, $j = 1, \ldots, q$, simultaneously)

$$\sum_{i=1}^{n} \frac{\partial \log p_k(\boldsymbol{\beta}_k \mid y_{ik}, z_i)}{\partial \boldsymbol{\beta}_k} = \sum_{i=1}^{n} (y_{ik} - \exp(\boldsymbol{\beta}_k^T \boldsymbol{w}_i)) \boldsymbol{w}_i - n\gamma \boldsymbol{D}^{-1} (\boldsymbol{\beta}_k - \boldsymbol{\delta}) = 0. \tag{3.5}$$

Upon applying a first order Taylor expansion of $\exp(\boldsymbol{\beta}_k^T \boldsymbol{w}_i)$ about a fixed $\boldsymbol{\beta}_k$, say $\tilde{\boldsymbol{\beta}}_k$, the estimation equation becomes approximately

$$\sum_{i=1}^{n} \frac{\partial \log p_k(\boldsymbol{\beta}_k \mid y_{ik}, z_i)}{\partial \boldsymbol{\beta}_k} = (\boldsymbol{Y} - \tilde{\boldsymbol{D}}_{1k}\tilde{\boldsymbol{\lambda}}_k)^T \boldsymbol{W} - \boldsymbol{\beta}_k^T (\boldsymbol{W}^T \tilde{\boldsymbol{D}}_{2k} \boldsymbol{W} + n\gamma \boldsymbol{D}^{-1}) + n\gamma \boldsymbol{\delta}^T \boldsymbol{D}^{-1} = 0, \tag{3.6}$$

where $\boldsymbol{W}$ is the $n \times q$ matrix with rows $\boldsymbol{w}_i^T$ $(i = 1, \ldots, n)$, and

$$\tilde{\boldsymbol{\lambda}}_k = \exp(\boldsymbol{W}\tilde{\boldsymbol{\beta}}_k) \quad \tilde{\boldsymbol{D}}_{1k} = \mathsf{Diag}(1 - \tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i) \quad \tilde{\boldsymbol{D}}_{2k} = \mathsf{Diag}(\exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i)). \tag{3.7}$$

A proof is provided in A.2 .

Given a $\tilde{\boldsymbol{\beta}}_k$ for the calculation of $\tilde{\boldsymbol{\lambda}}_k$, $\tilde{\boldsymbol{D}}_{1k}$ and $\tilde{\boldsymbol{D}}_{2k}$, Equation (3.6) results in

$$\hat{\boldsymbol{\beta}}_k = \left(\boldsymbol{W}^T \tilde{\boldsymbol{D}}_{2k} \boldsymbol{W} + n\gamma \boldsymbol{D}^{-1}\right)^{-1} \left[(\boldsymbol{Y} - \tilde{\boldsymbol{D}}_{1k}\tilde{\boldsymbol{\lambda}}_k)^T \boldsymbol{W} + n\gamma \boldsymbol{\delta}^T \boldsymbol{D}^{-1}\right]. \tag{3.8}$$

The $\boldsymbol{\beta}_k$ parameters can thus be estimated by iteratively calculating (3.7) and (3.8); the algorithm can be initialised by choosing $\tilde{\boldsymbol{\beta}}_k$ to be the not-penalised MLE.

3.2.3 Estimation of the Environmental Gradient

The LLR criterion (3.1) is now defined in terms of the posteriors,

$$\mathsf{LLR}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{k=1}^{s} \left(\log p_k(y_{ik} \mid z_i, \boldsymbol{\beta}_k) + \log g(\boldsymbol{\beta}_k) - \log p(y_{ik} \mid z_i, \boldsymbol{\beta})\right). \tag{3.9}$$

The LLR has to be maximised in terms of $\boldsymbol{\alpha}$, but note that the terms originating from the priors do not depend on $\boldsymbol{\alpha}$. Hence, (3.9) reduces to (3.1).

In A.2.1 we give details on a convenient iterative estimation algorithm based on Fisher scoring.

Often more than one environmental gradient is required to understand the species-environment relationship. We propose to work along the lines of **?** and **?**. The coefficients of the first environmental gradient, say $\boldsymbol{\alpha}_1$ is obtained by maximising the LLR criterion (3.9) and it results in the environmental scores $z_{1i} = \boldsymbol{\alpha}_1^t \boldsymbol{x_i}$ $(i = 1, \ldots, n)$.

The second environmental score must provide new information, unrelated to the first dimension. To this end, the environmental variables $\boldsymbol{x}_i$ are regressed on the scores of the first dimension. In particular, the p regression models, $x_{ij} = \zeta_{0j} + \zeta_{1j} z_{1i} + \epsilon_{ij}$ $(j = 1, \ldots, p)$, with $\mathrm{E}(\epsilon_{ij}) = 0$ are fitted using ordinary least squares. The resulting residuals, $e_{ij} = x_{ij} - \hat{\zeta}_{0j} - \hat{\zeta}_{1j} z_{1i}$, are known to be uncorrelated with the regressor (environmental score) of the first dimension. The matrix $\boldsymbol{X}$ is now replaced with the matrix $\boldsymbol{E}$ with rows $\boldsymbol{e}_i^T = (e_{i1}, \ldots, e_{ip})$ and this matrix serves as the new environmental matrix for obtaining the second environmental gradient. More gradients can be found by repeating this procedure (regressing $\boldsymbol{x}$ on all environmental scores).

3.3 Simulation Study

3.3.1 Simulation Setup

In this simulation study we evaluate the new method empirically.

Data will be simulated for species with bell-shaped response functions and for species with U-shaped response functions. Parameters are chosen such that the U-shaped

functions can be better separated than the bell-shaped functions, along some gradient. Classical methods are thus expected to find this gradient, whereas the BECOA method is designed to find another gradient along which mostly the bell-shaped functions are well separated.

We proceed with the following steps for the generation of the simulated data: (1) construct an environmental data matrix, $\boldsymbol{X}$, with observations of $p = 4$ environmental variables measured on $n = 44$ sampling locations; (2) specify two environmental gradients ($\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$); (3) specify $s = 20$ bell-shaped and U-shaped species response functions along the environmental gradients; (4) simulate 44 abundances for each of the 20 species. Details follow.

1. The 44×4 matrix $\boldsymbol{X}$ is formed by the four principle components of a part of the environmental data matrix used in Section 3.4. The measurements of calcium (Ca^{2+}), magnesium (Mg^{2+}) and potassium (K^{+}) and Silicate were considered.

2. The environmental gradients are orthonormal and set to $\boldsymbol{\alpha}_1^T = (1/\sqrt{2}, -1/\sqrt{2}, 0, 0)$ and $\boldsymbol{\alpha}_2^T = (0, 0, 1/\sqrt{2}, -1/\sqrt{2})$. The 44 corresponding environmental scores on the first two dimensions are calculated as $\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{\alpha}_1$ and $\boldsymbol{z}_2 = \boldsymbol{X}\boldsymbol{\alpha}_2$.

3. The response functions of the first twelve species are bell-shaped β-functions. In particular, for species $k = 1, \ldots, 12$,

$$\mathrm{E}\,(Y|z) = f_k^{\text{bell}}(z) = s_k(z-a)^{\eta_k}(b-z)^{\zeta_k} \quad a \le z \le b, \tag{3.10}$$

where for scale parameter $s_k > 0$ the expected abundance is always positive and it reaches its minimal value 0 at the two boundaries a and b. The function reaches

its maximum at $z = \frac{a\zeta_k + b\eta_k}{\eta_k + \zeta_k}$ (optimum); this maximum equals $\frac{(\eta_k(b-a))^{\eta_k}(\zeta_k(b-a))^{\zeta_k}}{(\eta_k+\zeta_k)^{\eta_k+\zeta_k}}$.

Table 3.1 shows the parameter settings for the first twelve species.

Table 3.1: The parameters used for the bell-shaped response functions for species $k = 1, \ldots, 12$. For all species, the scaling parameters s_k are set to $(\eta_k + \zeta_k)$ so as to make the maxima comparable.

	Species k											
	1	2	3	4	5	6	7	8	9	10	11	12
η_k	0.50	0.53	0.55	0.58	0.61	0.64	0.66	0.69	0.72	0.75	0.77	0.80
ζ_k	0.80	0.77	0.75	0.72	0.69	0.66	0.64	0.61	0.58	0.55	0.53	0.50
s_k	1.30	1.30	1.30	1.30	1.30	1.30	1.30	1.30	1.30	1.30	1.30	1.30
optimum	-0.24	-0.17	-0.10	-0.02	0.05	0.12	0.19	0.27	0.34	0.41	0.48	0.56

The remaining eight species show U-shaped response functions. In particular, for species $k = 13, \ldots, 20$,

$$\mathrm{E}\,(Y|z) = f_k^{\mathsf{U}}(z) = s_k(c - (z-a)^{\eta}(b-z)^{\zeta_k}) \quad a \leq z \leq b, \tag{3.11}$$

where $c = \frac{(\eta_k(b-a))^{\eta_k}(\zeta_k(b-a))^{\zeta_k}}{(\eta_k+\zeta_k)^{\eta_k+\beta_k}}$ is a location parameter to ensure positive abundances for all $z \in [0, 1]$. Table 3.2 shows the parameter settings for the eight species with U-shaped response functions.

Table 3.2: The parameters used for the U-shaped response functions for species $k = 13, \ldots, 20$. For all species, the scaling parameters s_k are set to $\frac{\eta_k+\zeta_k}{2}$ so as to make the maxima comparable.

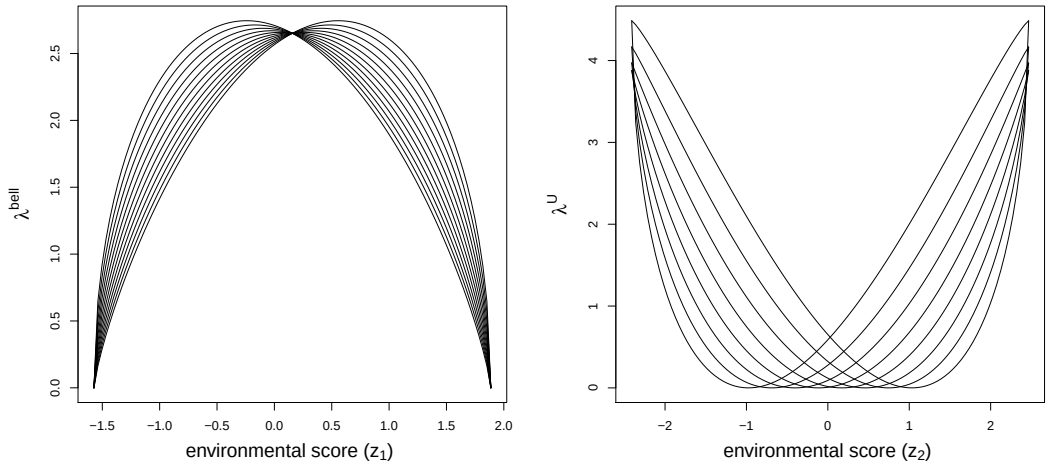
	Species k							
	13	14	15	16	17	18	19	20
η_k	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20
ζ_k	1.20	1.10	1.00	0.90	0.80	0.70	0.60	0.50
s_k	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
optimum	-0.98	-0.69	-0.40	-0.11	0.17	0.46	0.75	1.03

4. The abundances of the first twelve species are randomly generated from a Poisson distribution with mean parameter set to $f_k^{\mathsf{bell}}(z_{1i})$ with z_{1i} the environmental

scores along the first gradient ($k = 1, \ldots, 12$). The abundances of the other eight species use mean parameter $f_k^{\mathsf{U}}(z_{2i})$ with z_{2i} the environmental scores along the second gradient ($k = 13, \ldots, 20$).

The response functions are depicted in Figure 3.2. The parameters were chosen so that the U-shaped response functions are more separated than the bell-shaped response functions. Hence, the most important gradient found by the FCOA is more likely to be the second gradient, whereas the new method is expected to detect the first gradient.

Figure 3.2: The bell-shaped response functions of the first twelve species (left panel) and the U-shaped response functions of the next eight species (right panel)



Thousand datasets have been generated according to this procedure, and for each dataset the CCA, FCOA, and the new BECOA methods have been applied. Only the first environmental gradient is evaluated in the simulation study. Later, in the example section the use of the second dimension will be illustrated.

3.3.2 Results

Algorithms 1 and 2 are applied to one of the generate data to gain insight in the computational time. For the algorithm 1, it takes up to 7.41s CPU time, whereas, for the algorithm 2, the consumed CPU time is 12.05s. For each of the thousand generated datasets, the penalization parameter δ is set to vary from 0 to -1 in steps of -0.02. Figure 3.3(a) shows how the penalty affects the coefficients as estimated by the new BECOA method. As the penalization becomes stronger (i.e. moving from $\delta = 0$ to larger negative δ's), the estimates start deviating until they become stable for $\delta < -0.6$. The nervous behaviour of the coefficient estimates for $-0.6 < \delta < 0$ agrees with the large variance of these estimates (see intervals at the top of Figure 3.3(a)).

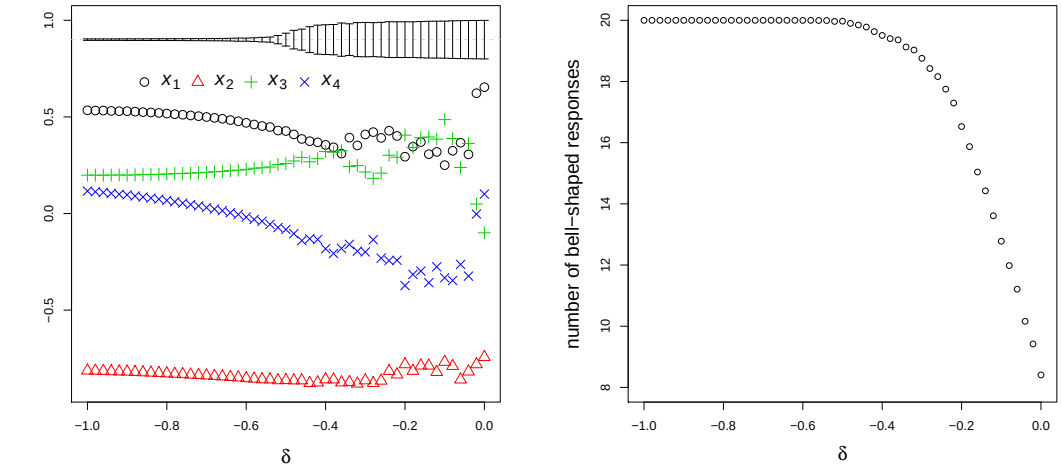
Figure 3.3(b) demonstrates that the new estimation method succeeds in increasing the number of species with bell-shaped response functions. Without penalisation only about eight species have bell-shaped response functions on average, but for $\delta < -0.5$ all 20 species do so in almost all simulations. Figure 3.3(c) shows how the $\hat{\beta}_{2k}$ are affected by the penalisation. Having forced all species to have bell-shaped response functions is not necessarily good. Therefore we have also assessed the goodness-of-fit with the total sum of squared errors (SSE) calculated from the fits of all 20 response functions: the effect of the penalisation on the goodness of fit is illustrated in Figure 3.3(d). It shows that the fit of the response functions of species 1-12 (with genuine bell-shaped response functions) slightly improves with increasing penalty, whereas for the other eight species the penalisation has a negative effect.

Figure 3.4 provides two diagnostic graphs that may be used for choosing an appropriate value for the penalisation parameter δ. The goal of our method is to find the gradient

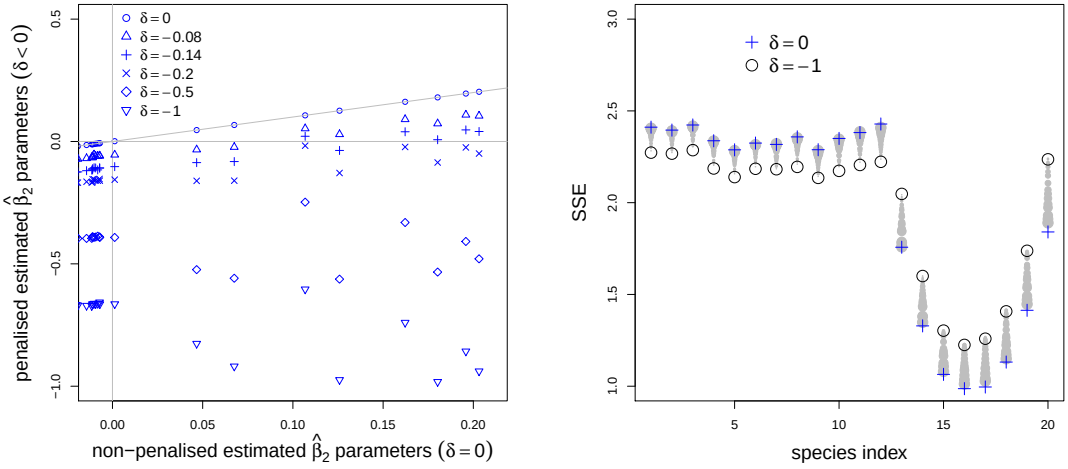
along which the species abundance distributions are maximally separated as measured by the LLR. However, only the contributions made by species with bell-shaped response functions should be included, because no ecologically meaningful interpretation will be given to the other species. We therefore define the average LLR (aLLR) as the LLR of Equation (3.9), but excluding the species with U-shaped fitted response functions, and divided by the number of bell-shaped fitted response functions. The left panel of Figure 3.4 shows a graph of the relative change of aLLR plotted against the penalty parameter. The relative change is computed as $\frac{\text{average LLR}(\delta) - \text{average LLR}(\delta=0)}{\text{average LLR}(\delta=0)}$. Hence, a minus sign in the percent change indicates the separation of the bell-shaped response functions at δ is worse than at $\delta = 0$. The construction of the right panel of Figure 3.4 is similar, but showing the average SSE (aSSE) as a measure for the quality of fit of the bell-shaped response functions. For this simulation study, the left panel of Figure 3.4 suggests that for $\delta > -0.6$ the penalisation has hardly a negative effect on the separation of the bell-shaped response functions, whereas the right panel indicates that the penalisation has a positive effect on the quality of the fit for species with bell-shaped response functions.

We conclude that, for the settings of the simulation study, the method works well. For $\delta = -0.6$, almost all species can have bell-shaped response functions, with overall a good goodness-of-fit, while loosing only about 10% of separation between the response functions.

Figure 3.3: Results of the simulation study: (a) the averages of the estimated environmental gradients as a function of the penalty parameter δ; the intervals shown on top are proportional to the total variance of the estimates. (b) the average number of bell-shaped response functions as a function of the penalty parameter δ. (c) for each of the 20 species the graph shows the evolution of the $\hat{\beta}_{2k}$'s as δ changes. (d) for each of the 20 species the graph shows the evolution of the Sum of Squared Errors (SSE) of the fits of the response functions for the penalty parameter moving from $\delta = 0$ (symbol: +) to $\delta = -1$ (symbol: O); the dots represent the intermediate results with larger dots representing smaller penalisation.

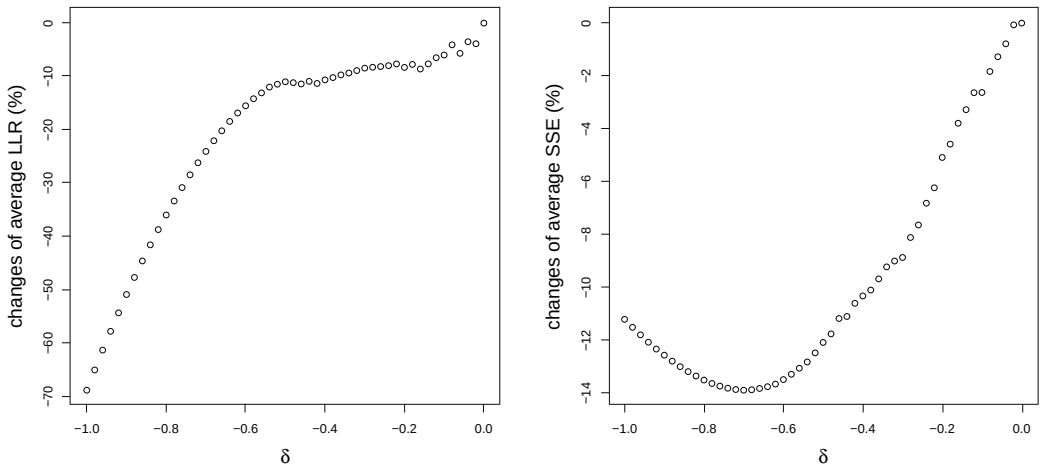


(a) (b)



(c) (d)

Figure 3.4: The relative changes of average LLR (left) and average SSE (right) as a function of the penalty parameter δ.



3.4 Example

We consider data from a limnology study conducted by **?**. They collected 45 water samples from lakes in ice-free regions along the east Antarctic coastline. Within the water samples seven physical and chemical characteristics were measured (environmental variables), including the concentrations of major ions (NH_4^+, K^+, Mg^{2+}, Ca^{2+} and Cl^-), silicate and the dissolved organic carbon (DOC). More than 500 microbial species were identified using a Massive Parallel Sequencing technique (Roche 454) and their relative abundances were calculated. Here we only use data from the species with more than three nonzero records, resulting in a reduction to 199 species. Each of the environmental variables was standardised to zero mean and unit variance prior to analysis. The data are analysed with three methods: BECOA, CCA and FCOA. For the latter method we considered a quadratic log-linear Poisson regression model as in Equation (3.3). All calculations are performed with **?**; for CCA the *vegan* R package

has been used.

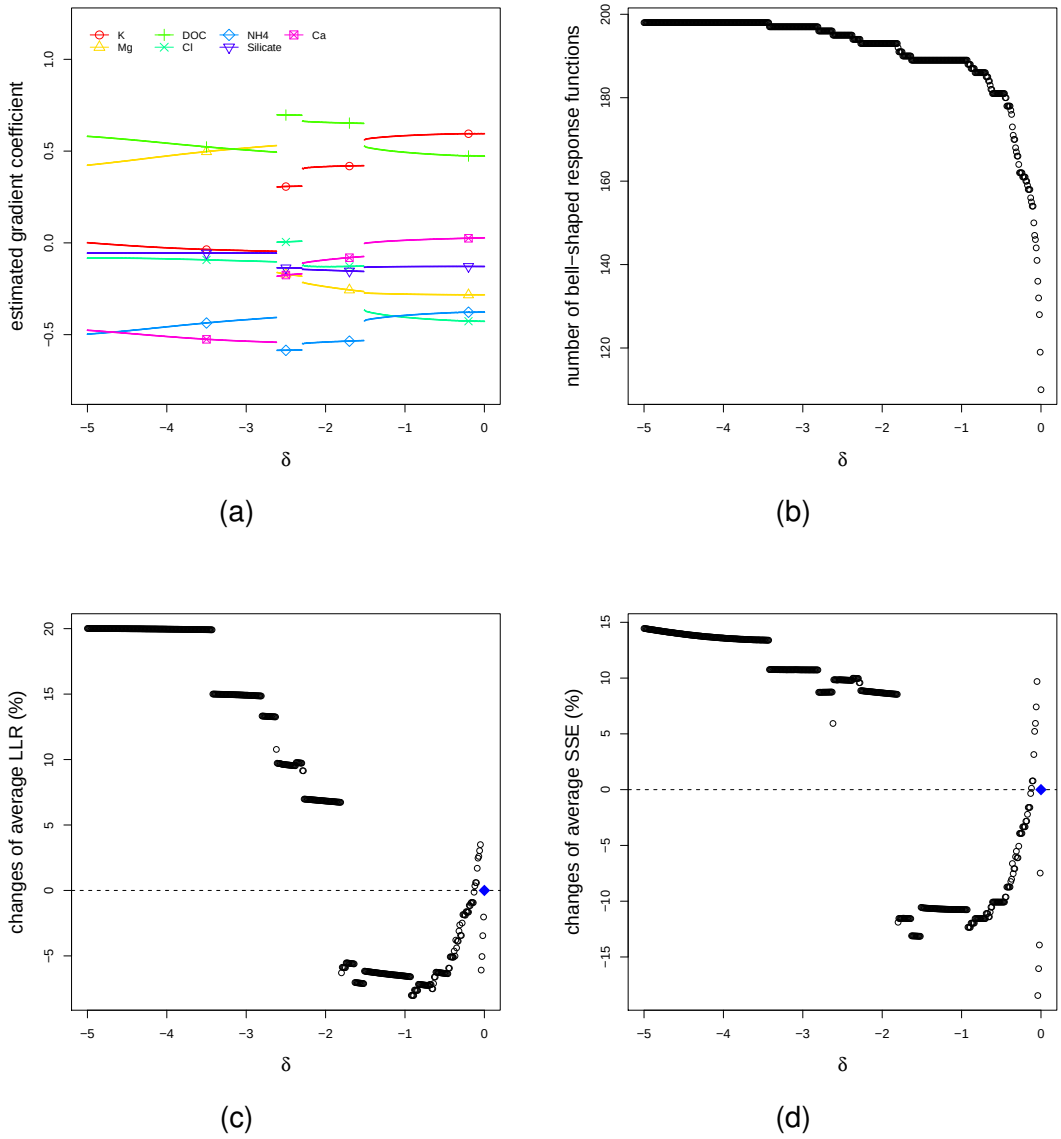
3.4.1 Selection of the tuning parameter

The normalised estimated coefficient of the environmental gradient for the first dimension are presented in Figure 3.5(a) as a function of the δ penalty parameter. All coefficients are affected by δ and the adjustment is sometimes quite substantial (e.g. the sign of coefficient of Mg^{2+} changed from negative to positive). Figure 3.5 (b) demonstrates a large increase in the number of species with bell-shaped response functions as the penalisation becomes heavier. At about $\delta = -3.5$ this number reaches 199. From Figure 3.5 (c) we see that the average LLR climbs with increasing penalisation, i.e. harder penalisation leads to more discrimination between the species with bell-shaped response functions. Figure 3.5 (d) shows the effect of penalisation on the average SSE. From this graph we conclude that, among the species with bell-shaped response functions, the aSSE quickly increases with penalisation, then ($-2 < \delta < -0.2$) it decreases, before ($\delta < -2$) increasing again.

Figure 3.6 shows the aLLR and aSSE, but now computed by 10-fold cross-validation so as to give a more honest assessment. See A.2.3 for details. From this graph we conclude that a compromise may be obtained with $\delta \in [-2.5; -1]$, resulting in an increase in the number of bell-shaped response functions from 110 to 190 with minimal cost in terms of quality of fit and discrimination.

Figure 3.7 shows the result for the second dimension. Figure 3.7(a) reveals that additional to the important environmental variables displayed in Figure 3.5(a), Silicate plays an important role in the second most discriminating direction. Along the second gra-

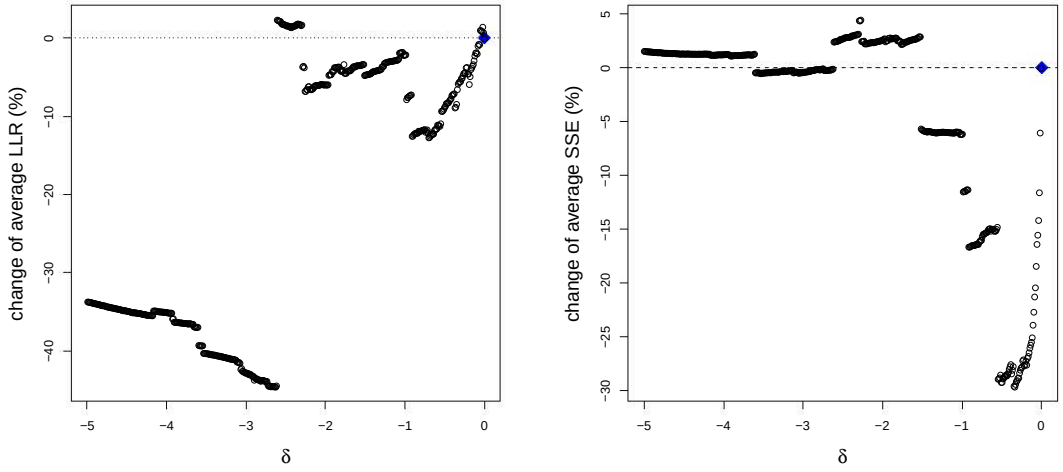
Figure 3.5: Results for the case study in the first dimension. Estimated coefficients of environmental gradient (a) and the average number of bell-shaped response functions (b) as a function of penalty parameter δ. Relative changes of average LLR (c) and average SSE (d) as a function of penalty parameter δ.



dient we can obtain at most 170 species with bell-shaped response functions (Figure 3.7(b)) at the cost of about 5% reduction in the aLLR (Figure 3.7(c)). The aSSE, as shown in Figure 3.7(d), falls down to 20% in almost a linear way.

The cross-validated aLLR and aSSE (Figure 3.8) display similar descending trends. This brings us to an appropriate δ of about -0.7, resulting in no cost in terms of sep-

Figure 3.6: Cross-validated (10-fold) results for the case study in the first dimension. Relative change of the average LLR (a) and average SSE (b) as a function of the penalty parameter δ.



(a) (b)

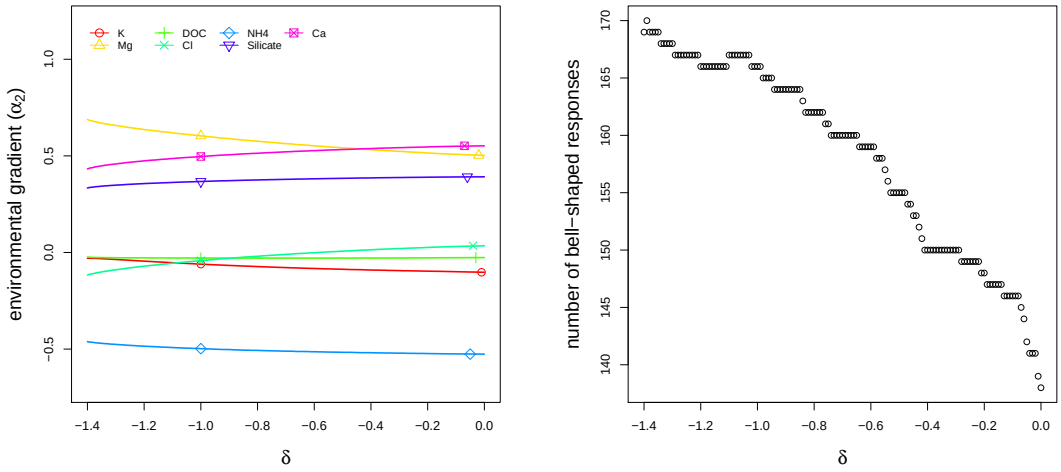
aration of the bell-shaped response functions and an increase of aSSE of about 10%, while more than 20 additional species are modelled with bell-shaped response functions.

3.4.2 Discussion

The estimated coefficients from three different ordination methods are listed in Table 3.3. The new method clearly has triggered changes in the estimates. Based on the discussion from the previous section, we have selected $\delta = -1.7$ and $\delta = -0.7$ for dimensions 1 and 2, respectively.

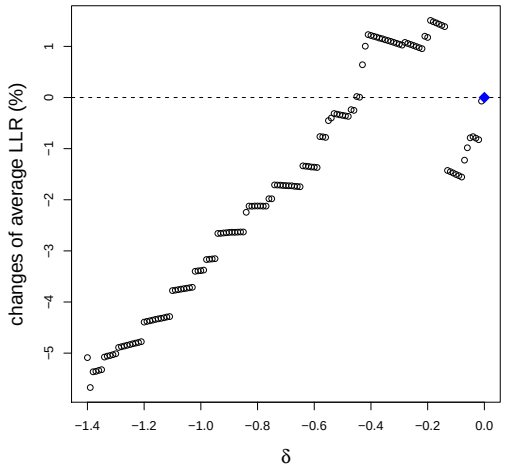
For the first environmental gradient direction we conclude that FCOA is not too dissimilar from the new method, whereas CCA gives quite different results. Table 3.3 also gives a closer insight into the consequences of the differences. The apparently small difference between BECOA and FCOA corresponds to a more than doubling of the

Figure 3.7: Results for the case study in the second dimension. Estimated coefficients of environmental gradient (a) and the average number of bell-shaped response functions (b) as a function of penalty parameter δ. Relative changes of average LLR (c) and average SSE (d) as a function of penalty parameter δ.

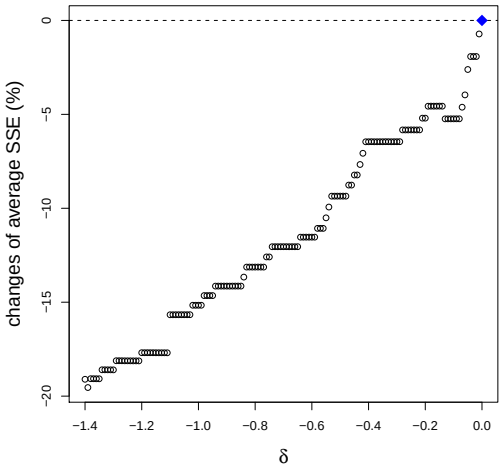


(a)

(b)



(c)

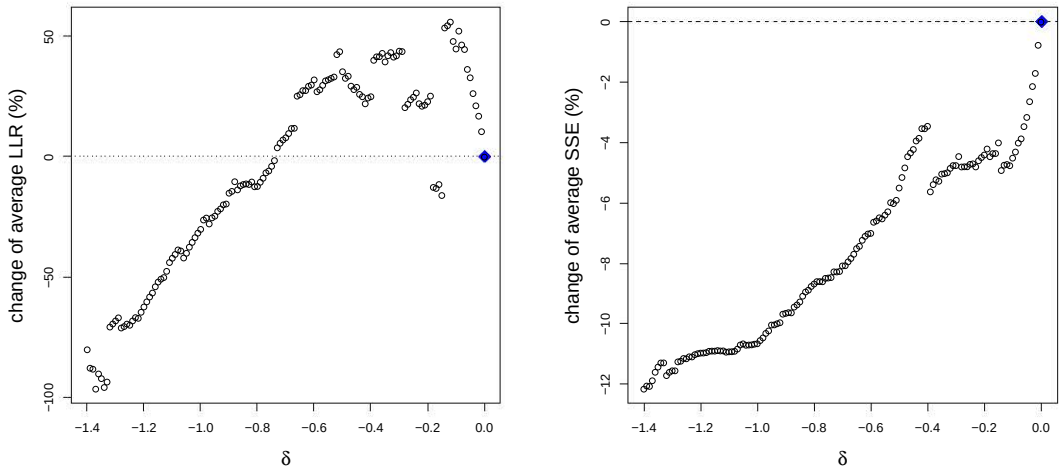


(d)

number of bell-shaped response functions, in favour of BECOA. Moreover, among the bell-shaped response functions, the quality of the fit of those from BECOA is better. By construction CCA fits bell-shaped response functions to all species, but the average MSE is worse than with BECOA.

Along the second dimension all three methods give different solutions. Table 3.3 shows

Figure 3.8: Cross-validated (10-fold) results for the second dimension. Relative change of the average LLR (a) and average SSE (b) as a function of the penalty parameter δ.



(a) (b)

that BECOA succeeds in fitting 166 bell-shaped response functions, whereas FCOA has only 66. As for the first dimension, BECOA gives the smallest average MSE among the bell-shaped response functions.

In A.2.2 results of the joint model fits for the first and the second dimension are given for all three ordination methods. Again BECOA appears on top in terms of MSE among the bell-shaped response functions, while FCOA has the poorest performance.

3.4.3 Ordination diagram

The results from the previous section are graphically presented as an ordination graph, similar as in **?**; see Figure 3.9. Information about the species optima, sampling sites and environmental gradients can be read from the graph. We argue that the data-analyst should only look at species for which a bell-shaped response function is obtained. Hence, only the 159 species for which in both ordination dimensions bell-

Table 3.3: Comparison of the estimated environmental gradients and the model fits from three ordination methods applied to the Antarctic lakes data. Dimension 1 and Dimension 2 refer to models fitted with the environmental scores on dimensions 1 and 2, respectively. MSE gives the mean squared error calculated only among Bell-shaped species, MSE^* stands for the mean squared error calculated from all species.

	Dimension 1			Dimension 2		
	BECOA	FCOA	CCA	BECOA	FCOA	CCA
number	190	80	199	166	66	199
$\sqrt{MSE}$	127.57	181.76	161.88	129.76	168.75	161.71
$\sqrt{MSE^*}$	156.38	143.94	161.88	152.73	157.16	161.71

Estimated environmental gradient

	Dimension 1			Dimension 2		
	BECOA	FCOA	CCA	BECOA	FCOA	CCA
K^+	0.4186	0.4409	0.1939	-0.0783	0.6754	0.2814
Mg^{2+}	-0.2565	-0.1866	-0.0275	0.5640	0.1272	-0.0399
DOC	0.6525	0.6056	-0.3890	-0.0291	-0.2000	-0.5647
Cl^-	-0.1280	-0.3797	-0.0181	-0.0098	0.5804	-0.0262
NH_4^+	-0.5355	-0.4861	0.4581	-0.5108	-0.1743	0.6650
Silicate	-0.1536	-0.1534	-0.0236	0.3784	0.2051	-0.0343
Ca^{2+}	-0.0808	-0.0073	-0.2723	0.5209	-0.2799	-0.3953

shaped response functions were found, are plotted. The plotting symbols indicate in what dimensions their response functions were U or bell-shaped if no penalisation were applied ($\delta = 0$).

This allows for deducing species-site relationships and the importance of environmental variables on species abundance distributions. The ordination diagrams of CCA and FCOA are included as Figure 3.10 and 3.11

Due to the large number of species the readability of Figure 3.9 is poor. In order to increase the interpretability, an ordination diagram with only ten randomly-selected species is presented in Figure 3.12. We conclude that particularly NH_4^+, K^+ and DOC determine the first dimension, while Mg^{2+}, Ca^{2+} and NH_4^+ dominate the second dimension.

Figure 3.9: Ordination diagram of the BECOA analysis of the Antarctic lake data, with penalisation parameter δ being -1.7 for the first dimension and =0.7 for the second dimension. Numbers represent lakes. The points represent the species optima, with symbols indicating the shape of the corresponding species response function when $\delta = 0$: p1, U-shaped in 1st and 2nd dimension; p2, bell-shaped in 1st dimension; p3, bell-shaped in 2nd dimension; p4, bell-shaped in 1st and 2nd dimension.

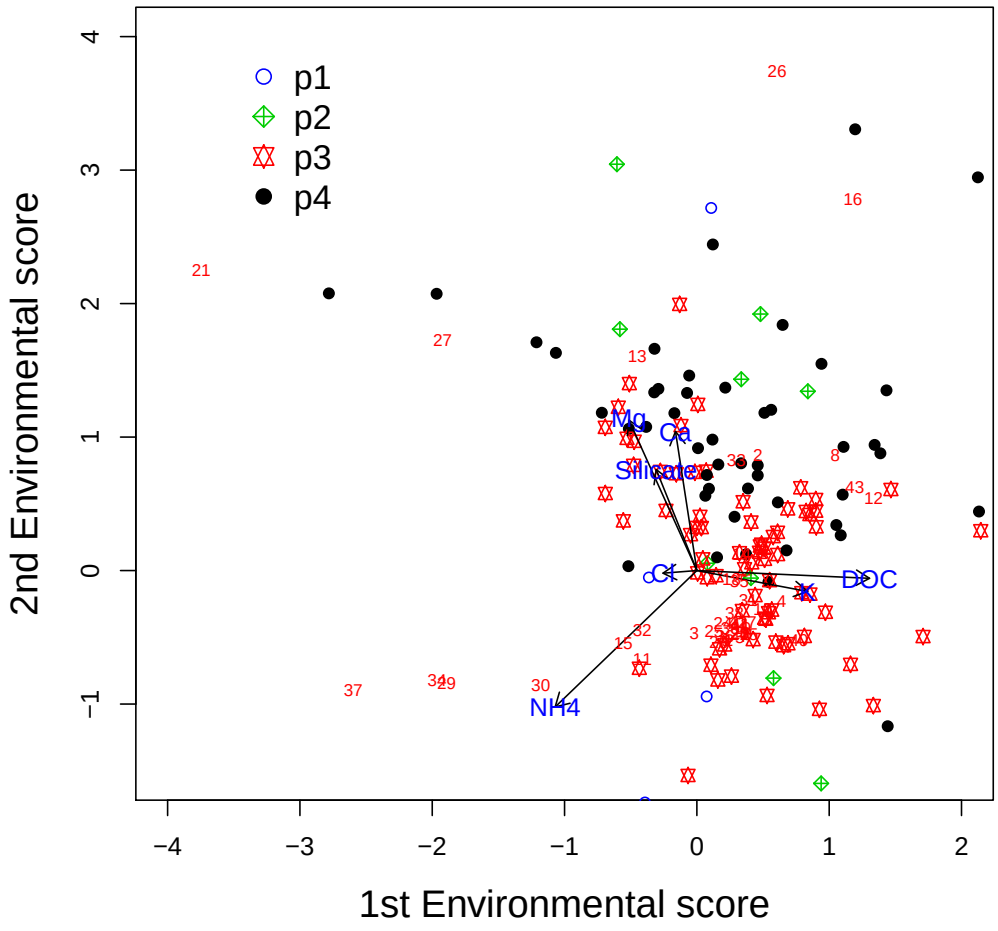


Figure 3.10: Ordination diagram of the CCA analysis of a subset of Antarctic data.

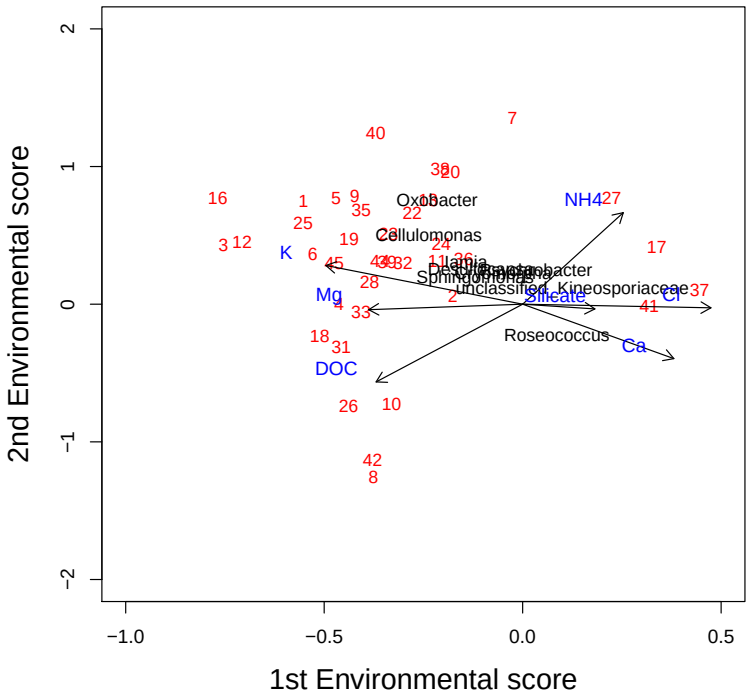


Figure 3.11: Ordination diagram of the FCOA analysis of a subset of Antarctic data.

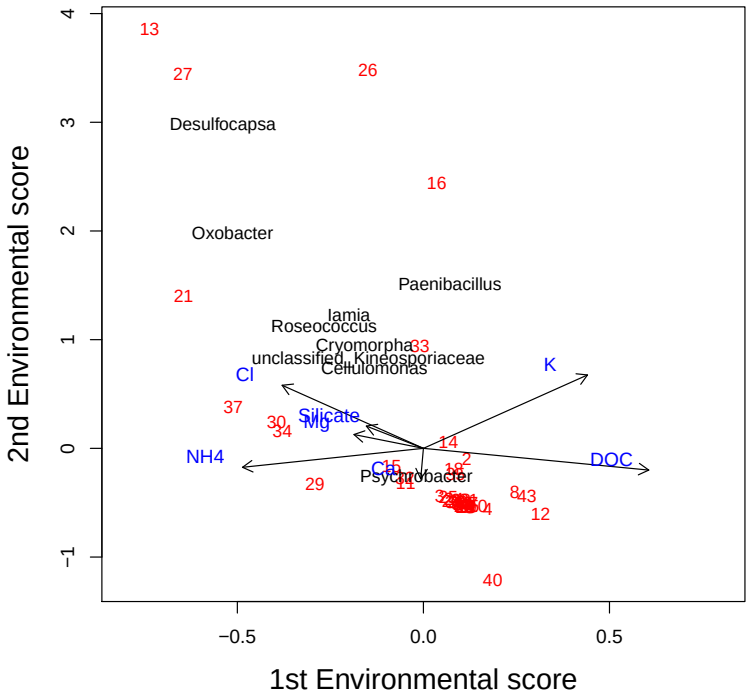
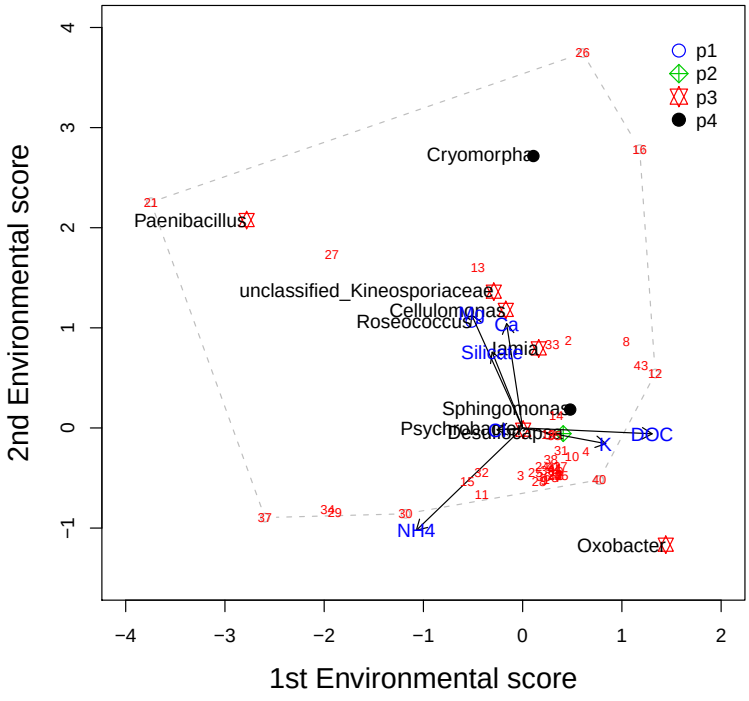


Figure 3.12: Ordination diagram of the BECOA analysis of a subset of the Antarctic lake data, with penalisation parameter δ being -1.7 for the first dimension and =0.7 for the second dimension. Numbers represent lakes. The points represent the species optima, with symbols indicating the shape of the corresponding species response function when $\delta = 0$: p1, U-shaped in 1st and 2nd dimension; p2, bell-shaped in 1st dimension; p3, bell-shaped in 2nd dimension; p4, bell-shaped in 1st and 2nd dimension. Species labels are added.



For species *Iamia*, for instance, the graph suggests that this species is most abundant at locations 33 and 2. This is confirmed by the observed abundances of this species: 105 and 74 at sites 33 and 2, respectively.

The results of the CCA analysis can also be depicted in an ordination graph. However, instead we summarise the differences between CCA and BECOA in a few statistics. We calculated the Euclidean distances between species optima and sites scores for the ordination diagrams produced by CCA and the new method. For each species we calculated the log-ratios of the distance from CCA over the distance from BECOA. A log-ratio close to zero indicates that both methods give about the same conclusion with respect to the preference of that species for that particular site. In the first dimension the median is 0.11 and the first and third quartiles are -0.94 and 1.36. Thus half of the species-location pairs have distance-ratio's more extreme than 0.11 and 23. For the second dimension, the median is 0.78 and the quartiles are -0.10 and 1.65, or half of the species-location pairs have distance-ratio's more extreme than 0.80 and 45.

3.5 Conclusions

Constrained ordination analysis (COA) methods aim at finding dimensions in the environmental space along which the species abundance response functions are maximally separated, allowing for explaining differences between the species environmental niches. Results from such analyses are typically graphically displayed as a biplot, which shows dots for species optimal environmental conditions. However, some COA methods do not enforce bell-shapes, but they do provide parameter estimates used for the construction of the biplots, which, consequently, are misleading and conclusions

are prone to error. Other methods do enforce the bell-shape, but by doing so they may result in poor fits of the response functions.

In this paper we have proposed a COA that searches for environmental gradients along with *bell-shaped* response functions are well separated. This is accomplished by setting up a tailored penalised maximum likelihood method that penalises U-shaped response functions. As a result environmental gradients can be identified that especially separate the ecologically meaningful bell-shaped response functions. We advise to remove species from the biplot for which no bell-shaped response functions are found for the first two environmental gradients.

In particular, two algorithms were proposed. The first method gives a simple implementation allowing for the use of Poisson regression routines available in most software packages. The second procedure requires the implementation of a variant of iteratively reweighted least squares, but it gives faster estimation times.

In Appendix B of **?** a computationally efficient algorithm is proposed to find a least square estimator subject to both nonnegative and sum constraints based on the lasso-path algorithm [**?**]. Here, we do not necessarily need the sum constraint because sparseness will not further improve the interpretability of the ordination analysis, but it may be worthwhile to further investigate the method for imposing nonnegative (or, equivalently, nonpositive) parameter estimates and study how this can be extended to a likelihood setting. Due to the use of the flexible LLR criterion, the BECOA can be extended to accommodate non-parametrically modelled response function following **?** or **?**. The latter proposed using B-splines and an asymmetric penalty for enforcing unimodal response functions.

Through a simulation study and a case study from a metagenomics aquatic limnology study, we have demonstrated the added value of our methods. For choosing an appropriate value of the penalisation tuning parameter, we have proposed a few diagnostic graphs related to the quality of the fit of the response functions and to the extent of separation of the bell-shaped response functions.

Although we only focused on modelling the species abundances, our method can be easily adapted to absence/presence data. This can be accomplished by replacing the Poisson regression with logistic regression (details are given in A.2.4). Similar adaptations may be considered when overdispersion (negative binomial regression) or an excess of zero abundances (zero inflated Poisson or zero inflated negative binomial regression) is expected.

Chapter 4

Semiparametric Gini index model

Summary: Biodiversity is defined as the variation of the living organisms within a defined space, this space can be as specific as a pond ecosystem or as general as the whole planet. The Gini index can be used to quantify this biodiversity. In this chapter, we first show the link between the Gini index quantity to the first two L-moments. Next, a semiparametric model is presented together with an appropriate approximation of the form of the link function. The proposed model allows us to model the Gini index as a function of covariates. The statistical inference on the model parameters is established and empirically evaluated in simulation studies. The method is also illustrated on a case study.

Key words: Biodiversity, Gini index, L2 moment, influence function, regression models.

4.1 Introduction

Biodiversity is the measure of the variety of an environment and it is a key concept to applications in traditional macroecology. For instance, in forest management, biodiversity of a forest is the indicator for sustainable use of the wood resources. It is well known that despite their small size, microorganisms have a huge impact on other ecosystem. Nevertheless studies of microbial biodiversity have been restricted due to the limitation of the traditional culture-dependent method for microbial species identification. In the last decade, the emergence of the more efficient and low-cost modern genome sequencing techniques, especially the next-generation sequencing technologies, have given a boost to research about microbial biodiversity in a variety of fields. Many of such studies result in a better understanding of the composition of the microbial community, and of their effect on the local ecosystem. For example, **?** and **?** have shown that shifts in the diversity of the human gut microbiome are linked to obesity and inflammatory bowel disease.

There are generally two measurements for describing the diversity: species richness and species evenness. Species richness gives the number of different species in a given habitat [**?**] and it is often considered [**?**]. Species evenness on the other hand, provides more insight into the composition of the community as compared to species richness. Species evenness ranges from 0 to 1, with larger values indicating that species are more equally abundant. If there is only one species that dominates a habitat, the evenness will be close to 0. If, on the other hand, multiple species occur in approximately the same amount, the evens will be close to 1. A species diversity index aims at combining the information in richness and evenness into a single index.

Several mathematical formulations have been brought up to quantify biodiversity [**???**], among which Gini index is appreciated due to its generality.

The Gini index, also known as the normalized Gini coefficient, has long been employed by economists to quantify the inequality among values of different levels of income or wealth. Its recognition in ecology as a measure of biodiversity has been discussed by **?**. The population estimator of the Gini index with interpretation based on a Lorenz curve has been shown in Section 1.3. **?** interpreted the Gini index as 'one half of the relative absolute mean abundance'. Consider a random sample of species abundance $\{Y_{ik} \mid i = 1, \ldots, n; k = 1, \ldots, K\}$. The Gini index estimate of location i, $\hat{G}_i$, is defined as

$$2\hat{G}_i = \frac{1}{\bar{Y}_i} \times \frac{\sum_{k=1}^{K} \sum_{l=1}^{K} |Y_{ik} - Y_{il}|}{K(K-1)}, \tag{4.1}$$

where $\bar{Y}_i$ is $\sum_{k=1}^{K} Y_{ik}/K$. The first term of the right-hand side of the equation is the inverse of the sample mean, whereas the second term is the sample mean of the absolute pairwise difference. In L-moment, these two terms are the sample estimator of the first (L1 moment) and second L-moment (L2 moment) respectively, see A.3.1 for an introduction of the L-moments. Hence Gini index can be estimated by half of the ratio of L2 moment to L1 moment. In Appendix A.3.2 a formal proof of the relationship of the population estimator of Gini index and the first two L-moments can be found.

Several biology scientists are in favour of the Gini index for measuring the biodiversity due to its generality [**???**]. The interpretation of the line of equality from Lorenz curve (Figure 1.6) is the scenario where the relative abundance of each species in the sample is equal. A graphical demonstration of perfect evenness is given in the left panel of Figure 4.1. The right panel, on the other hand, shows a community of species with

relatively high level of unevenness.

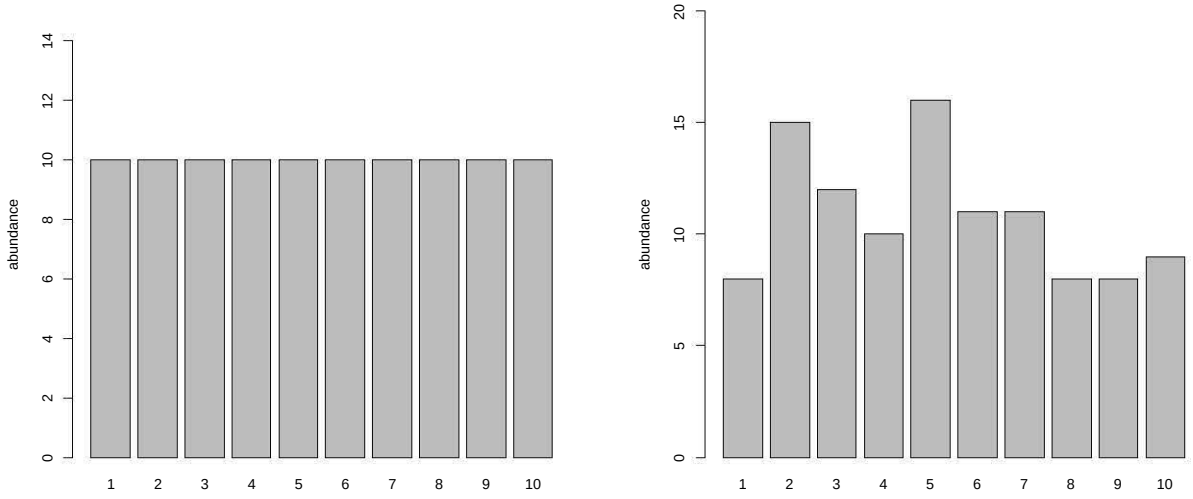


Figure 4.1: Plots of the scenarios of perfect evenness (left) and relative unevenness (right)

Given the sample estimator of the Gini index as in Equation (4.1), one can obtain an estimate of the Gini index of a community at each sampling location. Comparing the estimated Gini index across locations can help researcher to gain meaningful insights in the variation of the species diversity. In a regression framework, this is often done by regressing the estimated Gini index on several covariates through a general linear regression model. This simple approach is invalid in the sense that the estimated Gini indices are heterogeneous, i.e., the variance depends on the sampling location and this is against the assumption that general linear regression model requires. Instead of first estimating the Gini index and then using this estimate as the outcome in a general linear model, we model the Gini index directly as function of covariates.

This chapter is organised as follows: in Section 4.2 we develop a semiparametric regression model for the second L-moment. In Section 4.3 this model is empirically evaluated by a simulation study. We apply the new method to a real data set and show

how this second L-moment model can be used to construct a model for the Gini index in Section 4.4. In Section 4.5 we work out the estimation theory for the Gini index model. In Section 4.6 we empirically evaluate this model in a simulation study. In Section 4.7 we revisit the application for a final analysis. Several different aspects of the proposed model are discussed in Section 4.8.

4.2 L-moments model

In Section 4.2.1 we briefly review the restricted moment model for independent and identically distributed (i.i.d.) random observations. Since the abundance data are clustered according to location, we discuss the restricted moment model for clustered data in Section 4.2.2. This model will be the model to describe the L1-moment. In Section 4.2.3 the model on L2 moment is formally introduced.

4.2.1 Restricted moment model

Consider i.i.d. random observations $\{(Y_i, \boldsymbol{X}_i) \mid i = 1, \ldots, n\}$, then the restricted moment model [**??**] is given by:

$$g[\mathrm{E}(Y_i \mid \boldsymbol{X}_i)] = \boldsymbol{X}_i^T \boldsymbol{\alpha}, \tag{4.2}$$

where $g(\cdot)$ is the link function and $\boldsymbol{\alpha}$ is the vector of regression coefficients. A consistent estimator of $\boldsymbol{\alpha}$, assuming regularity and smoothness conditions, is obtained by solving

quasi-likelihood estimating equation [**?**]:

$$\sum_{i=1}^{n} \frac{\partial g^{-1}(\boldsymbol{X}_i^T \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} V^{-1}(\boldsymbol{X}_i)[Y_i - g(\boldsymbol{X}_i^T \boldsymbol{\alpha})] = \mathbf{0}, \tag{4.3}$$

where $V(\boldsymbol{X}_i) = (1/v)\text{var}(Y_i \mid \boldsymbol{X}_i)$.

4.2.2 L1 moment

The data subject to a biodiversity study often come from experiments where experimental units are determined before any data information is collected. Usually the experimental units are sampling locations. A number of chemical and physical environmental factors are measured at each sampling location and the records are stored in data matrix $\boldsymbol{X} = \{X_{ij}\}$, $i = 1, \ldots, n$ and $j = 1, \ldots, p$, here $\boldsymbol{X}_i$ represents the i^{th} row of $\boldsymbol{X}$ which contains all the measures recorded at sampling location i. The measured abundance are stored into matrix $\boldsymbol{Y}$, the i^{th} row of which refers to the observed abundance of total K species at sampling location i and the k^{th} column to the observed abundance of the k^{th} species over n locations. Consider clustered data $\{(\boldsymbol{Y}_i, \boldsymbol{X}_i) \mid i = 1, \ldots, n\}$, we model the L1 moment using generalized linear model as [**?**]

$$\mathrm{E}\left(Y_{ik} \mid \boldsymbol{X}_i\right) = g^{-1}(\boldsymbol{X}_i^T \boldsymbol{\alpha}), \tag{4.4}$$

where $k = 1, \ldots, K$. When Y_{ik} is assumed to follow a Poisson distribution, a logarithmic function is usually considered as the link function $g(\cdot)$. Model (4.4) is just a generalized linear model which is a special case of the restricted moment model (4.2) and the

regression coefficient $\boldsymbol{\alpha}$ can be estimated from the estimating equation:

$$\sum_{i=1}^{n} \frac{\partial g^{-1}(\boldsymbol{Z}_i\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^T} \left(\boldsymbol{A}_i^{1/2}\boldsymbol{R}_i\boldsymbol{A}_i^{1/2}\right)^{-1} [\boldsymbol{Y}_i - g^{-1}(\boldsymbol{Z}_i\boldsymbol{\alpha})] = \mathbf{0}, \tag{4.5}$$

where $\boldsymbol{Z}_i$ is $K \times p$ dimensional covariate matrix and $\boldsymbol{R}$ is a $p \times p$ symmetric matrix and serves as 'working' correlation matrix [**?**]. When the working correlation matrix is correctly specified the estimator of $\boldsymbol{\alpha}$ will be efficient. If the correlation matrix is misspecified, this efficiency property is lost, but the estimator is still consistent [**?**].

4.2.3 L2 moment

An L2 moment model has many similarities with so called *Probabilistic Index Models*. This section is therefore inspired by the work of **?** and **?**. We first work out the model for i.i.d. data and then show how it can be extended to the clustered setting.

Model formulation

In this section we propose a model that models the mean absolute difference of two randomly drawn variables as a function of covariates. Let $(Y, \boldsymbol{X})$ and $(Y', \boldsymbol{X}')$ be i.i.d. random observations, where Y denotes the univariate outcome and $\boldsymbol{X}$ the d-dimensional regressor. The L2 moment model is formulated as:

$$\mathrm{E}(|Y - Y'| \mid \boldsymbol{X}, \boldsymbol{X}') = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}), \tag{4.6}$$

where $m(\cdot)$ is a non-negative function and $\boldsymbol{\beta}$ is the d-dimensional parameter vector. Since $\mathrm{E}(|Y - Y'| \mid \boldsymbol{X}, \boldsymbol{X}') = \mathrm{E}(|Y' - Y| \mid \boldsymbol{X}', \boldsymbol{X})$, the function $m(\cdot)$ must be symmetric,

i.e., $m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}) = m(\boldsymbol{X}', \boldsymbol{X}; \boldsymbol{\beta})$. When the symmetry condition is not fulfilled, model (4.6) can be still appropriate when an ordering relation is imposed on the covariates. The *lexicographic ordering* is an example of such an ordering, see **?** and **?** for more details about lexicographic ordering.

For notational simplicity, we will refer to model (4.6) as the *L2 moment model*. To make a distinction between the models for which the symmetry condition holds and model for which specific restriction is set, we define the covariates set $\mathcal{X}$ with elements $(\boldsymbol{X}, \boldsymbol{X}')$ for which model (4.6) is defined. For covariates without restriction, notation $\mathcal{X}_0$ is used. In summary, the L2 moment model is defined as:

$$\mathrm{E}(|Y - Y'| \mid \boldsymbol{X}, \boldsymbol{X}') = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}), \quad (\boldsymbol{X}, \boldsymbol{X}') \in \mathcal{X}. \tag{4.7}$$

Model (4.7) imposes restriction on the conditional distribution of Y given $\boldsymbol{X}$ but without fully specifying the distribution, and is therefore a semiparametric model. We further impose the function $m(\cdot)$ to be related to a linear predictor, say $\boldsymbol{Z}^T\boldsymbol{\beta}$ where $\boldsymbol{Z}$ is a known function of $\boldsymbol{X}$ and $\boldsymbol{X}'$, via

$$m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}) = g^{-1}(\boldsymbol{Z}^T\boldsymbol{\beta}), \tag{4.8}$$

where $g(\cdot)$ is a sufficiently smooth function that maps $[0, \infty[$ on to the range of $\boldsymbol{Z}^T\boldsymbol{\beta}$. In most of the cases, the range of $\boldsymbol{Z}^T\boldsymbol{\beta}$ is just the real line. In the next paragraph we examine what choices for $g(\cdot)$ and $\boldsymbol{Z}$ are sensible.

Model approximation

In the study of biodiversity, data include the species abundance of a target set of species at each sampling location together with several measures of the covariates. Typically the abundance of a species is assumed to follow a Poisson or negative binomial distribution [**?**]. In order to find a feasible form of function $m(\cdot)$ in model (4.7), we need to work out the expectation of the absolute difference of Y and Y'. **?** has deduced the expression of the mean difference and the mean deviation for some positive integer valued discrete distributions. Here we present the result relevant for model (4.7).

Let Y and Y' to denote two arbitrary independent positive integer valued random variables with $P_i = P\{Y = i\}$ and $P'_i = P\{Y' = i\}$ respectively, following **?**, it can be shown that:

$$\begin{aligned}
\mathrm{E}\left(|Y - Y'|\right) &= \sum_{i=0}^{\infty}\sum_{k=0}^{\infty} kP\{Y' - Y = k \mid Y = i\}P\{Y = i\} \\
&\quad + \sum_{i=0}^{\infty}\sum_{k=0}^{\infty} kP\{Y - Y' = k \mid Y' = i\}P\{Y' = i\} \\
&= \sum_{i=0}^{\infty}\sum_{k=0}^{\infty} kP_iP'_{i+k} + \sum_{i=0}^{\infty}\sum_{k=0}^{\infty} kP'_iP_{i+k}.
\end{aligned}$$

Assume now that Y and Y' are Poisson-distributed random variables with parameter λ and λ' so that

$$P_i = \frac{e^{-\lambda}\lambda^i}{i!}, \quad P'_i = \frac{e^{-\lambda'}\lambda'^i}{i!}.$$

The worked out expression of the mean of the absolute difference for two independent

Poisson random variables given by **?** is

$$\begin{aligned} \mathrm{E}\left(|Y-Y'|\right) &= e^{-\lambda-\lambda'}\left\{\sum_{i=0}^{\infty}\sum_{k=0}^{\infty}\frac{k(\lambda\lambda')^i\lambda'^k}{i!(i+k)!}+\sum_{i=0}^{\infty}\sum_{k=0}^{\infty}\frac{k(\lambda\lambda')^i\lambda^k}{i!(i+k)!}\right\} \\ &= e^{-\lambda-\lambda'}(A+B), \end{aligned} \tag{4.9}$$

with $A=\sum_{i=0}^{\infty}\sum_{k=0}^{\infty}\frac{k(\lambda\lambda')^i\lambda'^k}{i!(i+k)!}$ and $B=\sum_{i=0}^{\infty}\sum_{k=0}^{\infty}\frac{k(\lambda\lambda')^i\lambda^k}{i!(i+k)!}$. The mean of the absolute difference for two Possion random variables is now expressed by the exponential of the summation of the two Poisson parameters, multiplied by the sum $(A+B)$. Since the conditional mean of a Poisson distributed random response variable is typically modeled employing a log link, expression (4.9) implies that model (4.7) involves a double exponential. Since this double exponential complicates the interpretation of the model drastically, we propose in model (4.7) the simplified functional form

$$g^{-1}(\boldsymbol{Z}^T\boldsymbol{\beta})=\exp\left[(\boldsymbol{X}+\boldsymbol{X}')^T\boldsymbol{\beta}\right]. \tag{4.10}$$

Since this is merely an approximation, it is possible that this model is not appropriate in some settings since the term A and B may also depend on $\boldsymbol{X}$ and $\boldsymbol{X}'$ and since we replaced the double exponential by a single exponential. In Section 4.3 we therefore evaluate the model approximation (4.10) in a simulation study. In the following part of this section we demonstrate how the parameters in model (4.10) can be estimated and we work out the asymptotic distribution of the estimator.

Parameter estimation and statistical inference

Define $P_{ih} := |Y_i - Y_h|$ and $\boldsymbol{Z}_{ih} = \boldsymbol{X}_i + \boldsymbol{X}_h$ for each $(\boldsymbol{X}_i, \boldsymbol{X}_h) \in \mathcal{X}$ where $i, h = 1, 2, \ldots, n$. Model (4.10) can be re-expessed as a restricted moment model:

$$\mathrm{E}(P_{ih} \mid \boldsymbol{X}_i, \boldsymbol{X}_h) = \exp(\boldsymbol{Z}_{ih}^T \boldsymbol{\beta}) \quad (\boldsymbol{X}_i, \boldsymbol{X}_h) \in \mathcal{X}. \tag{4.11}$$

Here P_{ih} is referred to as the *pseudo-outcome*. Model (4.11) resembles a semiparametric restricted moment model [**?**], we therefore propose the following estimating equation for $\boldsymbol{\beta}$:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{(i,h)\in\mathcal{P}} \boldsymbol{A}(\boldsymbol{Z}_{ih}; \boldsymbol{\beta})\{P_{ih} - g^{-1}(\boldsymbol{Z}_{ih}^T \boldsymbol{\beta})\} = \boldsymbol{0}, \tag{4.12}$$

where $\mathcal{P}$ is the set of indices (i, h) for which $(\boldsymbol{X}_i, \boldsymbol{X}_h) \in \mathcal{X}$, and $\boldsymbol{A}(\boldsymbol{Z}_{ih}; \boldsymbol{\beta})$ is a p-dimensional vector of functions of the $\boldsymbol{X}_i$ and $\boldsymbol{X}_h$. We set

$$\boldsymbol{A}(\boldsymbol{Z}_{ih}; \boldsymbol{\beta}) = \frac{\partial g^{-1}(\boldsymbol{Z}_{ih}^T \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}. \tag{4.13}$$

Note that the pseudo-outcomes are not mutually independent. This correlation exists among the pseudo-outcomes which share a common outcome, for example, consider three independent outcomes Y_i, $i = 1, 2, 3$, then $|Y_1 - Y_2|$ is associated with $|Y_1 - Y_3|$, $|Y_3 - Y_1|$, $|Y_2 - Y_3|$, $|Y_3 - Y_2|$ and $|Y_2 - Y_1|$. This correlation structure is sometimes called *cross-correlation* [**?**].

Due to the explicit correlation structure of our model, the distributional properties of the estimator of the regression coefficients do not follow from the theory of generalized estimating equation [**?**]. However, Lemma 1 states that the dependency in the pseudo-outcomes resembles the sparse correlation structure proposed by **?**.

Definition 1 (Sparse correlation)**.** *For each pseudo-outcomes* P_{ih}*, with* $(i,h) \in \mathcal{P} = \{(i,h) \mid i \neq h$ *and* $i,h = 1,\ldots,n\}$*, we define a set of indices* S_{ih} *such that for* k,l*,* P_{ih} *and* P_{kl} *are independent if* $(i,h) \notin S_{kl}$ *and* $(k,l) \notin S_{ih}$*. We refer to data as sparsely correlated if we can choose set* S_{ih}*,* $(i,h) \in \mathcal{P}$ *such that* $Mm = O(|\mathcal{P}|)$ *where* $|\mathcal{P}|$ *is the number of pseudo-outcomes in the set,* M *is the maximal number of paris in* S_{ih} *and* m *is the size of the largest subset* $\mathbf{T}$ *of* $\mathcal{P}$ *such that* $(i,h) \notin S_{kl}$ *and* $k,l \notin S_{ij}$ *for all pairs* $(i,h),(k,l) \in \mathbf{T}$*.*

Note that $g^{-1}(\boldsymbol{Z}_{ih}^T\boldsymbol{\beta}) = g^{-1}(\boldsymbol{Z}_{hi}^T\boldsymbol{\beta})$ and the function $\boldsymbol{A}(\cdot)$, shown in (4.13), follows $\boldsymbol{A}(\boldsymbol{Z}_{ih};\boldsymbol{\beta}) = \boldsymbol{A}(\boldsymbol{Z}_{hi};\boldsymbol{\beta})$. Therefore

$$
\begin{aligned}
&\sum_{(i,h)} \boldsymbol{A}(\boldsymbol{Z}_{ij};\boldsymbol{\beta})\{P_{ih} - g^{-1}(\boldsymbol{Z}_{ih}^T\boldsymbol{\beta})\} = \mathbf{0} \\
\Leftrightarrow\ &\sum_{i=1}^{n-1}\sum_{h=i+1}^{n} \boldsymbol{A}(\boldsymbol{Z}_{ih};\boldsymbol{\beta})\{P_{ih} - g^{-1}(\boldsymbol{Z}_{ih}^T\boldsymbol{\beta})\} + \sum_{h=1}^{n-1}\sum_{i=h+1}^{n} \boldsymbol{A}(\boldsymbol{Z}_{ih};\boldsymbol{\beta})\{P_{ih} - g^{-1}(\boldsymbol{Z}_{ih}^T\boldsymbol{\beta})\} = \mathbf{0} \\
\Leftrightarrow\ &\sum_{i=1}^{n-1}\sum_{h=i+1}^{n} \boldsymbol{A}(\boldsymbol{Z}_{ih};\boldsymbol{\beta})\{P_{ih} - g^{-1}(\boldsymbol{Z}_{ih}^T\boldsymbol{\beta})\} = \mathbf{0}.
\end{aligned}
$$

Hence, when the symmetric condition holds, only $n(n-1)/2$ summations are required to solve the estimating equation.

Lemma 1. *The pseudo-outcomes,* $\{P_{ih} \mid (i,h) \in \mathcal{P}\}$ *and* $\mathcal{P} = \{(i,h) \mid i < h = 2,\ldots,n\}$*, possess the sparse correlation structure.*

We refer to **?** for the proof of Lemma 1. To investigate the asymptotic property of the parameter estimator, we need to define the true β-parameter, say β_0, as the unique

solution of

$$\lim_{n\to\infty} \mathrm{E}\left\{|\mathcal{P}|^{-1} \sum_{(i,h)\in\mathcal{P}} \boldsymbol{A}(\boldsymbol{Z}_{ih};\boldsymbol{\beta})\left[|Y_i - Y_h| - g^{-1}(\boldsymbol{Z}_{ih}^T\boldsymbol{\beta})\right]\right\} = \mathbf{0}. \tag{4.14}$$

Where $|\mathcal{P}|$ denotes the number of elements in $\mathcal{P}$, we consider the following assumptions:

A1 the pseudo-outcomes are sparsely correlated as in Lemma 1;

A2 the link function $g(\cdot)$ and the variance function $V(\cdot)$ have three continuous derivatives;

A3 the true parameter β_0, as defined by (4.14), is in the interior of a convex parameter space;

A4 there exist a vector $\boldsymbol{W}$ and positive definite matrix $\boldsymbol{T}$ such that

$$|\mathcal{P}|^{-1} \sum_{(i,h)\in\mathcal{P}} \boldsymbol{Z}_{ih} \xrightarrow{p} \boldsymbol{W} \text{ and } |\mathcal{P}|^{-1} \sum_{(i,h)\in\mathcal{P}} \boldsymbol{Z}_{ih}\boldsymbol{Z}_{ih}^T \xrightarrow{p} \boldsymbol{T};$$

A5 $\limsup n^{-1}\mathrm{Var}\sum_{(i,h)\in\mathcal{P}} P_{ih} > 0$.

Theorem 1. *Let $\hat{\beta}$ denote the solution of estimating equation (4.12). Under assumptions A1 to A5, as $n \to \infty$, $\sqrt{n}(\hat{\beta} - \beta_0)$ converges in distribution to a multivariate Gaussian distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.*

Theorem 2. *Let*

$$\boldsymbol{U}_{ih}(\boldsymbol{\beta}) = \boldsymbol{A}(\boldsymbol{Z}_{ih};\boldsymbol{\beta})\{P_{ih} - g^{-1}(\boldsymbol{Z}_{ih}^T\boldsymbol{\beta})\},$$

under assumptions A1 to A5, the variance-covariance matrix Σ can be consistently estimated by the sandwich estimator

$$n\hat{\boldsymbol{\Sigma}}_{\hat{\beta}} = \left(\sum_{(i,h)\in\mathcal{P}} \frac{\partial \boldsymbol{U}_{ih}(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^T} \right) \left(\sum_{(i,h)\in\mathcal{P}} \sum_{(k,l)\in\mathcal{P}} \phi_{ihkl} \boldsymbol{U}_{ih}(\hat{\boldsymbol{\beta}}) \boldsymbol{U}_{kl}^T(\hat{\boldsymbol{\beta}}) \right) \left(\sum_{(i,h)\in\mathcal{P}} \frac{\partial \boldsymbol{U}_{ih}(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \right)^{-1}, \tag{4.15}$$

where the indicator ϕ_{ihkl} is defined as $\phi_{ihkl} = 1$ if P_{ih} and P_{kl} are correlated and $\phi_{ihkl} = 0$ otherwise.

Theorem 1 and 2 follow directly from Theorem 7 of **?**.

To sum up, when the sample size n is large and Model (4.11) holds, $\hat{\beta}$ approximately follows a multivariate normal distribution with mean β_0 and variance-covariance matrix that can be estimated by $\hat{\boldsymbol{\Sigma}}_{\hat{\beta}}$.

The L2 moment model for clustered data

Recall the $\boldsymbol{Y} = \{Y_{ik}\}$ is used to denote the abundance matrix and $\boldsymbol{X} = \{X_{ij}\}$ the covariate matrix. Each $\boldsymbol{X}_i$, the environmental variables that characterise location i, is associated with K species abundance $\boldsymbol{Y}_i = \{Y_{i1}, \ldots, Y_{iK}\}$. The data are however not i.i.d., since that the sampling sites form clusters and within clusters the abundances are correlated. We therefore propose the following L2 moment model that accounts for the clustering.

$$\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid \boldsymbol{X}_i\right) = g^{-1}\left[(\boldsymbol{X}_i + \boldsymbol{X}_i)^T \boldsymbol{\beta}\right] \tag{4.16}$$

This model implies that the pseudo-outcomes are obtained by only comparing the abundances within sampling locations. In the following Lemma, we show that the pseudo-outcomes are still sparsely correlated.

Lemma 2. *The pseudo-outcomes,* $|Y_{ik} - Y_{il}|$ *with* $(k, l) \in \mathcal{P}$ *and* $\mathcal{P} = \{(k, l) \mid k < l = 2, \ldots, K.\}$*, possess the sparse correlation structure.*

Proof. Each $|Y_{ik} - Y_{il}|$ is correlated with $[K(K-1)-1]/2$ other pseudo-outcomes. Indeed $|Y_{ik} - Y_{il}|$ is correlated with $|Y_{id} - Y_{id'}|$, $d < d' = 2, \ldots, K$, $d \neq k$ and $d' \neq l$. That is the pseudo-outcomes that are associated with location i are correlated, thus $M = [K(K-1)-1]/2$. The largest set of pseudo-outcomes that are mutually independent is of size n. Then

$$Mn = \frac{n\left[(K-1)K-1\right]}{2} = \frac{nK^2 - nK - 2n}{2} = O(nK^2)$$

Since $O(|\mathcal{P}|) = O(\frac{nK(K-1)}{2}) = O(nK^2)$, the lemma holds. □

The pseudo-outcomes from clustered data are sparsely correlated and hence Theorem 2 holds.

4.3 Simulation study

In this section simulation studies are set up to assess three properties of model (4.11): (1) whether approximation (4.10) is appropriate for data following a Poission distribution; (2) to assess consistency of the estimators of the model parameters and (3) whether approximation (4.10) is appropriate when overdispersion is present.

4.3.1 Linear model

In this simulation we consider the univariate linear L2 moment model.

Simulation setting for i.i.d. data

We restrict the simulation study to the setting with a univariate predictor X. Multivariate predictors are beyond the scope of this dissertation. The predictor takes equally spaced values in the interval $[-1, 1]$ and the species abundance Y is generated from Poisson or negative binomial distribution with conditional expectation:

$$\mathrm{E}\left(Y_{ik} \mid X_i\right) = \lambda(X_i) = \exp\left(\alpha_0 + \alpha_1 X_i\right). \tag{4.17}$$

Note that X_i does not depend on k, since similar as in the case study we only consider a predictor that does not vary within a sampling location.

We fit the following L2 moment model:

$$\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i\right) = \exp\left(\beta_0 + \beta_1(X_i + X_i)\right) = \exp(\beta_0 + 2\beta_1 X_i), \tag{4.18}$$

where $k < l$. Note that we only compare outcomes within sampling location. We consider 3 data generating model for the abundance Y:

Setting 1: The response Y is generated from Poisson distribution with mean (4.17).

Setting 2: The response Y is generated from negative binomial distribution with mean (4.17) and dispersion parameter $\theta = 1$.

Setting 3: The response Y is generated from negative binomial distribution with mean (4.17)

and dispersion parameter $\theta = 10$.

Smaller dispersion parameter θ indicates data with larger dispersion. The number of species K is set to $K = 5, 10$ and 30. A total of 1000 Monte Carlo simulation runs are applied to investigate the distribution of the estimator of $\boldsymbol{\beta}$. All computations have been performed with the R software [**?**]. An overview of the simulation set-up is given in Table 4.1.

Table 4.1: Schematic overview of the simulation procedures.

1. Generate a series of n X_i's varying from -1 to 1 in steps of 0.1, thus $n = 21$.
2. Replicate each X_i K times.
3. The response $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iK})$ is generated from
   (a) setting 1: Poisson distribution with mean $\lambda(X_i) = \exp(\alpha_0 + \alpha_1 X_i)$.
   (b) setting 2: negative binomial distribution with mean $\lambda(X_i) = \exp(\alpha_0 + \alpha_1 X_i)$ and dispersion parameter $\theta = 1$
   (c) setting 3:negative binomial distribution with mean $\lambda(X_i) = \exp(\alpha_0 + \alpha_1 X_i)$ and dispersion parameter $\theta = 10$
4. Fit Model (4.18).

In Table 4.2 we proposed a procedure to obtaining unbiased estimator of the true β_0 and β_1.

Table 4.2: Schematic overview of obtaining the true regression coefficients β_0 and β_1, that do not depend on model (4.18).

1. Generate a series of X_i varying from -1 to 1 in steps of 0.1.
2. Replicate each X_i K times.
3. From model (4.18) it follows that

$$\beta_0 = \log\left[\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i = 0\right)\right]. \tag{4.19}$$

   To obtain β_0, we randomly generate 10^6 Poisson responses Y_{ik} and Y_{il} with mean $\lambda(X_i = 0) = \exp(\alpha_0)$, and approximate β_0 by using the empirical expectation in expression (4.19).
4. If we evaluate model (4.18) at $X_i = x_i$ and $X_i = x_i + 1$, it follows that

$$\log\left[\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i = x_i\right)\right] = \beta_0 + 2\beta_1 x_i,$$

   and

$$\log\left[\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i = x_i + 0.1\right)\right] = \beta_0 + 2\beta_1 x_i + 0.2\beta_1.$$

   Therefore

$$\beta_1(X_i = x_i) = \frac{\log\left[\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i = x_i + 0.1\right)\right] - \log\left[\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i = x_i\right)\right]}{0.2}. \tag{4.20}$$

   To obtain β_1 we randomly generated two 10^6 response vectors Y_{ik} and Y_{il} with mean $\lambda(X_i = x_i) = \exp\left(\alpha_0 + \alpha_1 x_i\right)$ and corresponding dispersion parameter θ. Next, we generate another two 10^6 vectors of abundance Y_{ik} and Y_{il} with mean $\lambda(X_i = x_i + 0.1) = \exp\left[\alpha_0 + \alpha_1(x_i + 0.1)\right]$ and corresponding dispersion parameter θ. We approximate β_1 by replacing the expectation in (4.20) by its empirical counterpart.

Simulation result

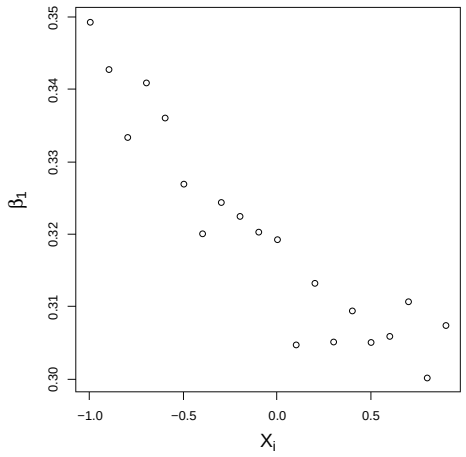
We set $\alpha_0 = \log(3)$, $\alpha_1 = \log(10/3)$ and the choice is made for mimicking the data from the case study in Section 4.4. Since model (4.18) is merely an approximation of the traditional L2-moment, it can be misspecified. Model (4.18) implies the following

$$\beta_0 = \log\left[\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i = 0\right)\right], \tag{4.21}$$

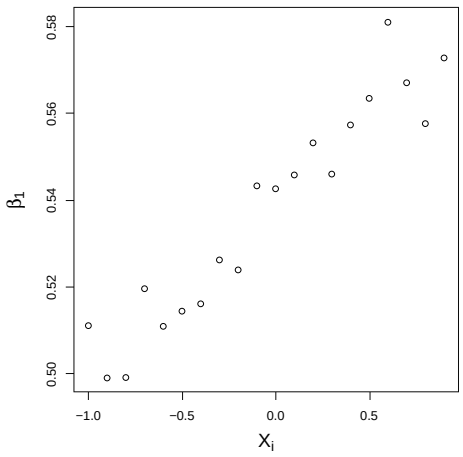
and

$$\beta_1 = \frac{\log\left[\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i = x_i + 0.1\right)\right] - \log\left[\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i = x_i\right)\right]}{0.2}. \tag{4.22}$$

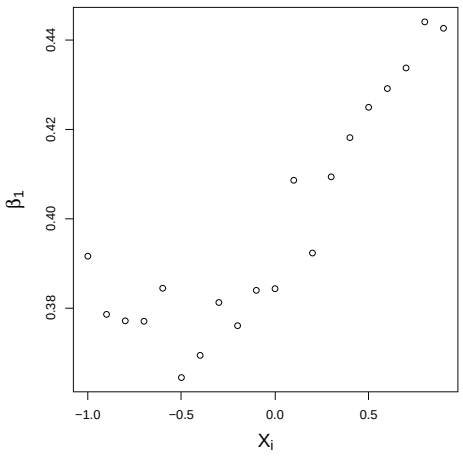
To obtain the true value of β_0 and β_1, we simulate 10^6 responses according to the three settings and use equations (4.21) and (4.22) to approximate β_0 and β_1. Because β_1 can actually depend on x_i, we write $\beta_1(x_i)$. If model (4.18) is appropriate, we however expect β_1 to be independent of x_i. Figure 4.2 roughly show linear relation of β_1 and the covariate X for all 3 settings. This indicates that model (4.18) is not a good approximation of the true model. However the dependence of β_1 on x_i is rather weak: all values of β_1 lie within 0.3 and 0.35. Tables 4.3, 4.4 and 4.5 show the simulation results according to the different settings. Since the true value of β depends slightly on x_i, we report the average of these values and denote the average by $\bar{\boldsymbol{\beta}}$. From the result, we conclude that in general, the empirical coverage gets improved when more replicates are present. The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are nearly unbiased, furthermore the variance-covariance matrices are unbiased as well. From table 4.4 and 4.5 we see that when K is large, the sandwich estimator overestimates the true variance, resulting in a conservative confidence interval. Although the results from Tables 4.3 - 4.5 supported the feasibility of model (4.18), we should not ignore the fact that the true value of β depends slightly on x_i from Figure 4.2. In the next section, we evaluate model (4.18) with an additional quadratic term of X.



(a)



(b)



(c)

Figure 4.2: Plot of the resulting β_1 based on the three simulation settings: (a) for setting 1, (b) for setting 2 and (c) for setting 3.

Table 4.3: Simulation result from setting 1, based on 1000 Monte Carlo runs. $\mathrm{Av}(\bar{\boldsymbol{\beta}})$ is the average of $\bar{\boldsymbol{\beta}}$; $\mathrm{Av}(\hat{\boldsymbol{\beta}})$ the average of $\hat{\boldsymbol{\beta}}$; $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ the empirical estimate of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$; $\mathrm{Av}(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$ the average of the estimate of the variance-covariance matrix using sandwich estimator and EC the empirical coverage of a 95% confidence interval.

K	$\mathrm{Av}(\bar{\boldsymbol{\beta}})$;	$\mathrm{Av}(\hat{\boldsymbol{\beta}})$	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\mathrm{Av}(\hat{\Sigma}_{\hat{\boldsymbol{\beta}}})$	EC(%)
5	$\begin{bmatrix} 0.65 \\ 0.32 \end{bmatrix}$	$\begin{bmatrix} 0.6327 \\ 0.3167 \end{bmatrix}$	$\begin{bmatrix} 0.0081 & -0.0009 \\ -0.0009 & 0.0056 \end{bmatrix}$	$\begin{bmatrix} 0.0078 & -0.0019 \\ -0.0019 & 0.0068 \end{bmatrix}$	$\begin{bmatrix} 92.3 \\ 92.8 \end{bmatrix}$
10	$\begin{bmatrix} 0.65 \\ 0.32 \end{bmatrix}$	$\begin{bmatrix} 0.6366 \\ 0.3172 \end{bmatrix}$	$\begin{bmatrix} 0.0032 & -0.0003 \\ -0.0003 & 0.0025 \end{bmatrix}$	$\begin{bmatrix} 0.0034 & -0.0007 \\ -0.0007 & 0.0029 \end{bmatrix}$	$\begin{bmatrix} 94.1 \\ 94.4 \end{bmatrix}$
30	$\begin{bmatrix} 0.65 \\ 0.32 \end{bmatrix}$	$\begin{bmatrix} 0.6403 \\ 0.3167 \end{bmatrix}$	$\begin{bmatrix} 0.0011 & -0.00007 \\ -0.00007 & 0.0007 \end{bmatrix}$	$\begin{bmatrix} 0.0010 & -0.0002 \\ -0.0002 & 0.0009 \end{bmatrix}$	$\begin{bmatrix} 94.1 \\ 96.6 \end{bmatrix}$

Table 4.4: Simulation result from setting 2, based on 1000 Monte Carlo runs. $\mathrm{Av}(\bar{\boldsymbol{\beta}})$ is the average of $\bar{\boldsymbol{\beta}}$; $\mathrm{Av}(\hat{\boldsymbol{\beta}})$ the average of $\hat{\boldsymbol{\beta}}$; $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ the empirical estimate of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$; $\mathrm{Av}(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$ the average of the estimate of the variance-covariance matrix using sandwich estimator and EC the empirical coverage of a 95% confidence interval.

K	$\mathrm{Av}(\bar{\boldsymbol{\beta}})$	$\mathrm{Av}(\hat{\boldsymbol{\beta}})$	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\mathrm{Av}(\hat{\Sigma}_{\hat{\boldsymbol{\beta}}})$	EC(%)
5	$\begin{bmatrix} 1.25 \\ 0.54 \end{bmatrix}$	$\begin{bmatrix} 1.2288 \\ 0.5443 \end{bmatrix}$	$\begin{bmatrix} 0.0164 & -0.0029 \\ -0.0029 & 0.0148 \end{bmatrix}$	$\begin{bmatrix} 0.0135 & -0.0086 \\ -0.0086 & 0.0142 \end{bmatrix}$	$\begin{bmatrix} 97.1 \\ 93.7 \end{bmatrix}$
10	$\begin{bmatrix} 1.25 \\ 0.54 \end{bmatrix}$	$\begin{bmatrix} 1.2363 \\ 0.5462 \end{bmatrix}$	$\begin{bmatrix} 0.0075 & -0.0008 \\ -0.0008 & 0.0061 \end{bmatrix}$	$\begin{bmatrix} 0.0034 & -0.0007 \\ -0.0007 & 0.0029 \end{bmatrix}$	$\begin{bmatrix} 97.5 \\ 95.4 \end{bmatrix}$
30	$\begin{bmatrix} 1.25 \\ 0.54 \end{bmatrix}$	$\begin{bmatrix} 1.2430 \\ 0.5444 \end{bmatrix}$	$\begin{bmatrix} 0.0022 & -0.0002 \\ -0.0002 & 0.0021 \end{bmatrix}$	$\begin{bmatrix} 0.0042 & -0.0027 \\ -0.0027 & 0.0049 \end{bmatrix}$	$\begin{bmatrix} 97.9 \\ 98.4 \end{bmatrix}$

Table 4.5: Simulation result from setting 3, based on 1000 Monte Carlo runs. $\mathrm{Av}(\bar{\boldsymbol{\beta}})$ is the average of $\bar{\boldsymbol{\beta}}$; $\mathrm{Av}(\hat{\boldsymbol{\beta}})$ the average of $\hat{\boldsymbol{\beta}}$; $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ the empirical estimate of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$; $\mathrm{Av}(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$ the average of the estimate of the variance-covariance matrix using sandwich estimator and EC the empirical coverage of a 95% confidence interval.

K	$\mathrm{Av}(\bar{\boldsymbol{\beta}})$	$\mathrm{Av}(\hat{\beta})$	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\mathrm{Av}(\hat{\Sigma}_{\hat{\boldsymbol{\beta}}})$	EC(%)
5	$\begin{bmatrix} 0.77 \\ 0.40 \end{bmatrix}$	$\begin{bmatrix} 0.7799 \\ 0.3974 \end{bmatrix}$	$\begin{bmatrix} 0.0078 & -0.0007 \\ -0.0007 & 0.0669 \end{bmatrix}$	$\begin{bmatrix} 0.0095 & -0.0037 \\ -0.0037 & 0.0097 \end{bmatrix}$	$\begin{bmatrix} 94.8 \\ 92.8 \end{bmatrix}$
10	$\begin{bmatrix} 0.77 \\ 0.40 \end{bmatrix}$	$\begin{bmatrix} 0.7813 \\ 0.3955 \end{bmatrix}$	$\begin{bmatrix} 0.0036 & -0.0003 \\ -0.0003 & 0.0029 \end{bmatrix}$	$\begin{bmatrix} 0.0041 & -0.0014 \\ -0.0014 & 0.0042 \end{bmatrix}$	$\begin{bmatrix} 96.2 \\ 94.9 \end{bmatrix}$
30	$\begin{bmatrix} 0.77 \\ 0.40 \end{bmatrix}$	$\begin{bmatrix} 0.7834 \\ 0.3968 \end{bmatrix}$	$\begin{bmatrix} 0.0010 & -0.00007 \\ -0.00007 & 0.0008 \end{bmatrix}$	$\begin{bmatrix} 0.0013 & -0.0004 \\ -0.0004 & 0.0014 \end{bmatrix}$	$\begin{bmatrix} 94.8 \\ 97.6 \end{bmatrix}$

4.3.2 Quadratic model

We consider the quadratic L2 moment model

$$\begin{aligned} \mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i\right) &= \exp\left(\beta_0 + \beta_1(X_i + X_i) + \beta_2(X_i^2 + X_i^2)\right) \\ &= \exp\left(\beta_0 + 2\beta_1 X_i + 2\beta_2 X_i^2\right), \end{aligned} \tag{4.23}$$

The simulation set-up in this part remains the same as in Table 4.1 but the model to fit is now model (4.23) in step 5. In the previous simulation study, the approach to obtaining β_0 (see Table 4.2) can be used for evaluating the β_0 for model (4.23), however the approach to obtaining β_1 is no longer applicable when the model is quadratic. In Table 4.6 it is explained how β_1 can be approximated.

Table 4.6: Schematic overview of obtaining the true regression coefficient β_1 that do not depend on model (4.23).

1. Generate a series of X_i varying from -1 to 1 in steps of 0.1.
2. Replicate each X_i K times.
3. If we evaluate model (4.23) at $X_i = x_i$, it follows that

$$\log\left[\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i = x_i\right)\right] = \beta_0 + 2\beta_1 x_i + 2\beta_2 x_i^2, \tag{4.24}$$

and we evaluate again model (4.23) at $X_i = -x_i$, it follows that

$$\log\left[\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i = -x_i\right)\right] = \beta_0 - 2\beta_1 x_i + 2\beta_2 x_i^2. \tag{4.25}$$

Therefore

$$\beta_1(X_i = x_i) = \frac{\log\left[\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i = x_i\right)\right] - \log\left[\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i = -x_i\right)\right]}{4x_i}. \tag{4.26}$$

To obtain β_1 we randomly generated two 10^6 response vectors Y_{ik} and Y_{il} with mean $\lambda(X_i = x_i) = \exp(\alpha_0 + \alpha_1 x_i)$ and the corresponding dispersion parameter θ. Next, we generate another two 10^6 vectors of abundance Y_{ik} and Y_{il} with mean $\lambda(X_i = -x_i) = \exp(\alpha_0 - \alpha_1 x_i)$ and the corresponding dispersion parameter θ. We approximate β_1 by replacing the expectation in (4.26) by its empirical counterpart.

Simulation result

Model (4.23) implies the following

$$\begin{aligned} x_i\beta_1 &= \frac{\log\left[\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i = x_i\right)\right] - \log\left[\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i = -x_i\right)\right]}{4} \\ &= Q(X_i). \end{aligned} \tag{4.27}$$

If model (4.23) is correctly specified, we expect that β_1 is independent of the value of X_i. However, we see that in the proposed estimating equation (4.26) for the true β_1,

X_i appears in the denominator and its value can be 0. To avoid numerical problem we check the trend of $x_i\beta_i$, i.e., $Q(X)$ over X. If β_1 does not depend on X, $Q(X)$ should show linear relation with X. If the relation is not linear it suggests misspecification of model (4.23), for example, a quadratic relation of $Q(X)$ and X may imply that β_1 still depends on X and thus a cubic term of X in the model might be necessary.

Figure 4.3 presents the relation of $Q(X)$ (equation (4.27)) versus X_i for the three settings. The solid line is the fitted line resulting from regressing $Q(X)$ against X. In all cases, we see that the points scattered closely around the solid line. This indicates that model (4.23) is a feasible approximation of the conditional L2 moment, and thus β_1 stays constant for different values of X.

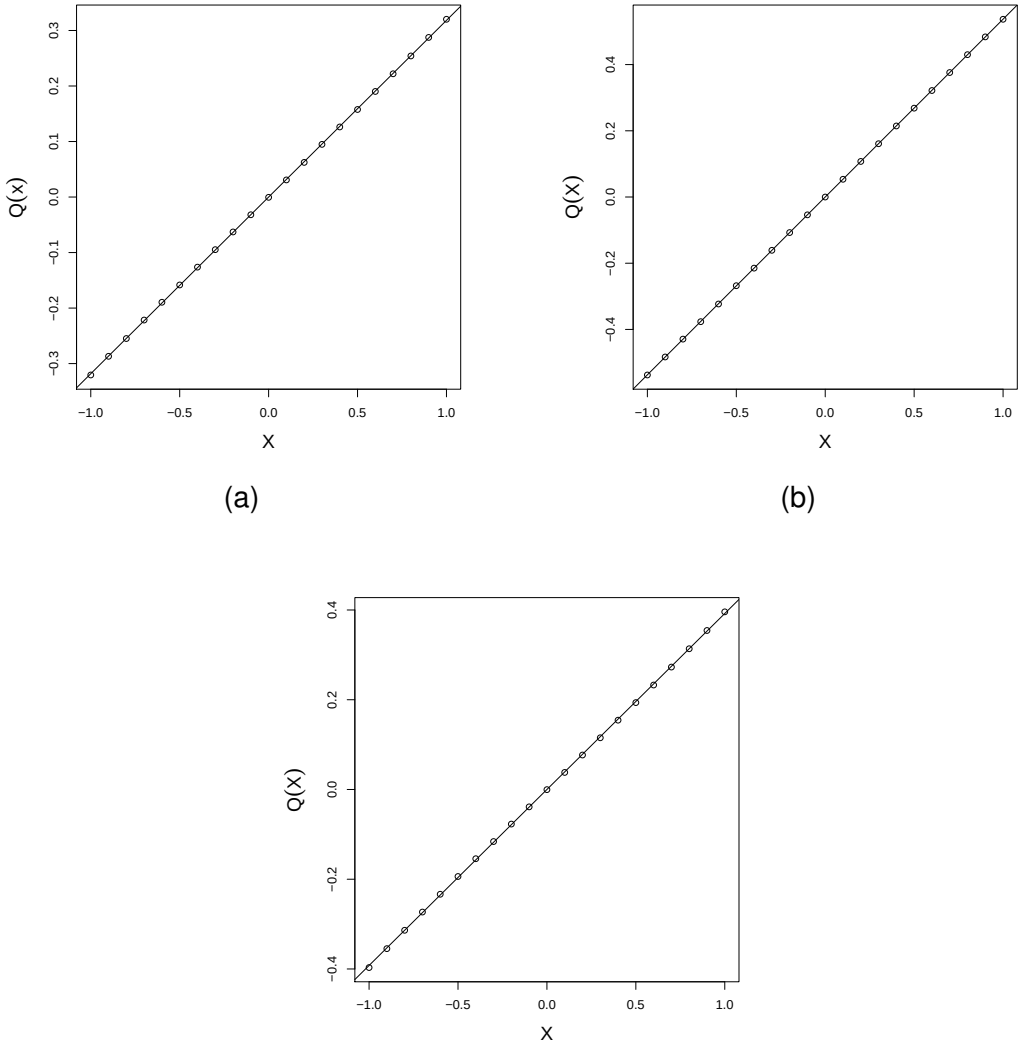


Figure 4.3: $Q(X)$, from equation (4.27), is plotted as a function of X for the three simulation settings: (a) for setting 1, (b) for setting 2 and (c) for setting 3. The solid line is the fitted line resulting from regressing $Q(X)$ against X.

Tables 4.7 - 4.9 are the simulation results from different settings. We use the slope of the fitted line in Figure 4.3 as the measure of the true β_1 and we denote this slop by $\bar{\beta}_1$. We see that including the quadratic term results in slightly better estimates and coverage of β_0 and β_1. However, the results from the linear settings are quite satisfying and the $\hat{\beta}_2$'s are relatively small (from -0.02 to 0.02) for all three setting, indicating that model (4.23) is appropriate for modelling the L2 moment. The estimators

of the regression coefficients obtained from solving the estimating equation (4.12) are consistent and the proposed model worked well when overdispersion is present.

Table 4.7: Simulation result for setting 1 of the quadratic model, based on 1000 Monte Carlo simulation runs. $\mathrm{Av}(\bar{\boldsymbol{\beta}})$ is the average $\bar{\boldsymbol{\beta}} = (\bar{\beta}_0, \bar{\beta}_1)^T$; $\mathrm{Av}(\hat{\boldsymbol{\beta}})$ the average of $\hat{\boldsymbol{\beta}}$; $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ the empirical estimate of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$; $\mathrm{Av}(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$ the estimate of the variance-covariance matrix using sandwich estimator and EC the empirical coverage of a 95% confidence interval for β_0 and β_1.

K	$\mathrm{Av}(\bar{\boldsymbol{\beta}})$	$\mathrm{Av}(\hat{\boldsymbol{\beta}})$	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\mathrm{Av}(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$	EC(%)
5	$\begin{bmatrix} 0.64 \\ 0.32 \end{bmatrix}$	$\begin{bmatrix} 0.6423 \\ 0.3181 \\ -0.0197 \end{bmatrix}$	$\begin{bmatrix} 0.01619 & 0.0002 & -0.0128 \\ 0.0002 & 0.0056 & -0.0011 \\ -0.0128 & -0.0011 & 0.018 \end{bmatrix}$	$\begin{bmatrix} 0.0182 & 0.0012 & -0.0158 \\ 0.0012 & 0.0055 & -0.0029 \\ -0.0158 & -0.0029 & 0.0219 \end{bmatrix}$	$\begin{bmatrix} 95.7 \\ 93.4 \end{bmatrix}$
10	$\begin{bmatrix} 0.64 \\ 0.32 \end{bmatrix}$	$\begin{bmatrix} 0.6445 \\ 0.3192 \\ -0.0144 \end{bmatrix}$	$\begin{bmatrix} 0.0082 & 0.0001 & -0.0063 \\ 0.0001 & 0.0024 & -0.0003 \\ -0.0063 & -0.0003 & 0.0082 \end{bmatrix}$	$\begin{bmatrix} 0.0081 & 0.0005 & -0.0070 \\ 0.0005 & 0.0023 & -0.0011 \\ -0.0070 & -0.0011 & 0.0098 \end{bmatrix}$	$\begin{bmatrix} 94.2 \\ 93.7 \end{bmatrix}$
30	$\begin{bmatrix} 0.64 \\ 0.32 \end{bmatrix}$	$\begin{bmatrix} 0.6456 \\ 0.3180 \\ -0.0097 \end{bmatrix}$	$\begin{bmatrix} 0.0022 & 0.0001 & -0.0017 \\ 0.0001 & 0.0007 & -0.0001 \\ -0.0017 & -0.0001 & 0.0024 \end{bmatrix}$	$\begin{bmatrix} 0.0025 & 0.0001 & -0.0022 \\ 0.0001 & 0.0007 & -0.0003 \\ -0.0022 & -0.0003 & 0.0031 \end{bmatrix}$	$\begin{bmatrix} 96.6 \\ 95.3 \end{bmatrix}$

Table 4.8: Simulation result for setting 1 of the quadratic model, based on 1000 Monte Carlo simulation runs. $\mathrm{Av}(\bar{\boldsymbol{\beta}})$ is the average $\bar{\boldsymbol{\beta}} = (\bar{\beta}_0, \bar{\beta}_1)^T$; $\mathrm{Av}(\hat{\boldsymbol{\beta}})$ the average of $\hat{\boldsymbol{\beta}}$; $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ the empirical estimate of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$; $\mathrm{Av}(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$ the estimate of the variance-covariance matrix using sandwich estimator and EC the empirical coverage of a 95% confidence interval for β_0 and β_1.

K	$\mathrm{Av}(\bar{\boldsymbol{\beta}})$	$\mathrm{Av}(\hat{\boldsymbol{\beta}})$	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\mathrm{Av}(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$	EC(%)
5	$\begin{bmatrix} 1.23 \\ 0.54 \end{bmatrix}$	$\begin{bmatrix} 1.2261 \\ 0.5396 \\ -0.0108 \end{bmatrix}$	$\begin{bmatrix} 0.0366 & 0.0008 & -0.0305 \\ 0.0008 & 0.0125 & -0.0030 \\ -0.0305 & -0.0030 & 0.0453 \end{bmatrix}$	$\begin{bmatrix} 0.0533 & 0.0109 & -0.0572 \\ 0.0109 & 0.0239 & -0.0270 \\ -0.0572 & -0.0270 & 0.0885 \end{bmatrix}$	$\begin{bmatrix} 95.5 \\ 95.5 \end{bmatrix}$
10	$\begin{bmatrix} 1.23 \\ 0.54 \end{bmatrix}$	$\begin{bmatrix} 1.230 \\ 0.5356 \\ 0.0165 \end{bmatrix}$	$\begin{bmatrix} 0.0191 & 0.0012 & -0.0163 \\ 0.0012 & 0.0053 & -0.0020 \\ -0.0163 & -0.0020 & 0.0225 \end{bmatrix}$	$\begin{bmatrix} 0.0270 & 0.0054 & -0.0300 \\ 0.0054 & 0.0093 & -0.0110 \\ -0.0300 & -0.0110 & 0.0456 \end{bmatrix}$	$\begin{bmatrix} 94.4 \\ 96.7 \end{bmatrix}$
30	$\begin{bmatrix} 1.23 \\ 0.54 \end{bmatrix}$	$\begin{bmatrix} 1.2334 \\ 0.5388 \\ 0.0151 \end{bmatrix}$	$\begin{bmatrix} 0.0060 & 0.0001 & -0.0051 \\ 0.0001 & 0.0017 & -0.0003 \\ -0.0051 & -0.0003 & 0.0073 \end{bmatrix}$	$\begin{bmatrix} 0.0079 & 0.0018 & -0.0087 \\ 0.0018 & 0.0029 & -0.0034 \\ -0.0087 & -0.0034 & 0.0137 \end{bmatrix}$	$\begin{bmatrix} 96.6 \\ 97.6 \end{bmatrix}$

Table 4.9: Simulation result for setting 1 of the quadratic model, based on 1000 Monte Carlo simulation runs. $\mathrm{Av}(\bar{\boldsymbol{\beta}})$ is the average $\bar{\boldsymbol{\beta}} = (\bar{\beta}_0, \bar{\beta}_1)^T$; $\mathrm{Av}(\hat{\boldsymbol{\beta}})$ the average of $\hat{\boldsymbol{\beta}}$; $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ the empirical estimate of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$; $\mathrm{Av}(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$ the estimate of the variance-covariance matrix using sandwich estimator and EC the empirical coverage of a 95% confidence interval for β_0 and β_1.

K	$\mathrm{Av}(\bar{\boldsymbol{\beta}})$	$\mathrm{Av}(\hat{\boldsymbol{\beta}})$	$\mathrm{Cov}(\hat{\beta})$	$\mathrm{Av}(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$	EC(%)
5	$\begin{bmatrix} 0.77 \\ 0.39 \end{bmatrix}$	$\begin{bmatrix} 0.7659 \\ 0.3915 \\ 0.0057 \end{bmatrix}$	$\begin{bmatrix} 0.0194 & 0.0005 & -0.0157 \\ 0.0005 & 0.0059 & -0.0010 \\ -0.0157 & -0.0010 & 0.0218 \end{bmatrix}$	$\begin{bmatrix} 0.0214 & 0.0021 & -0.0195 \\ 0.0021 & 0.0064 & -0.0044 \\ -0.0195 & -0.0044 & 0.0270 \end{bmatrix}$	$\begin{bmatrix} 95.1 \\ 94.7 \end{bmatrix}$
10	$\begin{bmatrix} 0.77 \\ 0.39 \end{bmatrix}$	$\begin{bmatrix} 0.7672 \\ 0.3925 \\ 0.0192 \end{bmatrix}$	$\begin{bmatrix} 0.0081 & 0.0001 & -0.0064 \\ 0.0001 & 0.0030 & -0.0005 \\ -0.0064 & -0.0005 & 0.0090 \end{bmatrix}$	$\begin{bmatrix} 0.0100 & 0.0009 & -0.0092 \\ 0.0009 & 0.0028 & -0.0018 \\ -0.0092 & -0.0018 & 0.0130 \end{bmatrix}$	$\begin{bmatrix} 96.3 \\ 93.4 \end{bmatrix}$
30	$\begin{bmatrix} 0.77 \\ 0.39 \end{bmatrix}$	$\begin{bmatrix} 0.7695 \\ 0.3923 \\ 0.0217 \end{bmatrix}$	$\begin{bmatrix} 0.0026 & 0.0001 & -0.0021 \\ 0.0001 & 0.0008 & -0.0002 \\ -0.0021 & -0.0002 & 0.0030 \end{bmatrix}$	$\begin{bmatrix} 0.0031 & 0.0003 & -0.0029 \\ 0.0003 & 0.0009 & -0.0005 \\ -0.0029 & -0.0005 & 0.0042 \end{bmatrix}$	$\begin{bmatrix} 96.9 \\ 95.5 \end{bmatrix}$

Figure 4.4 shows model diagnosis for model (4.23). The average of the fitted pseudo-outcomes based on model (4.23) are plotted against the average of the observed pseudo-outcomes from 1000 simulation runs. The predicted pseudo-outcomes correspond closely to the observed ones and this again indicates that model (4.23) is appropriate for modelling the L2 moment.

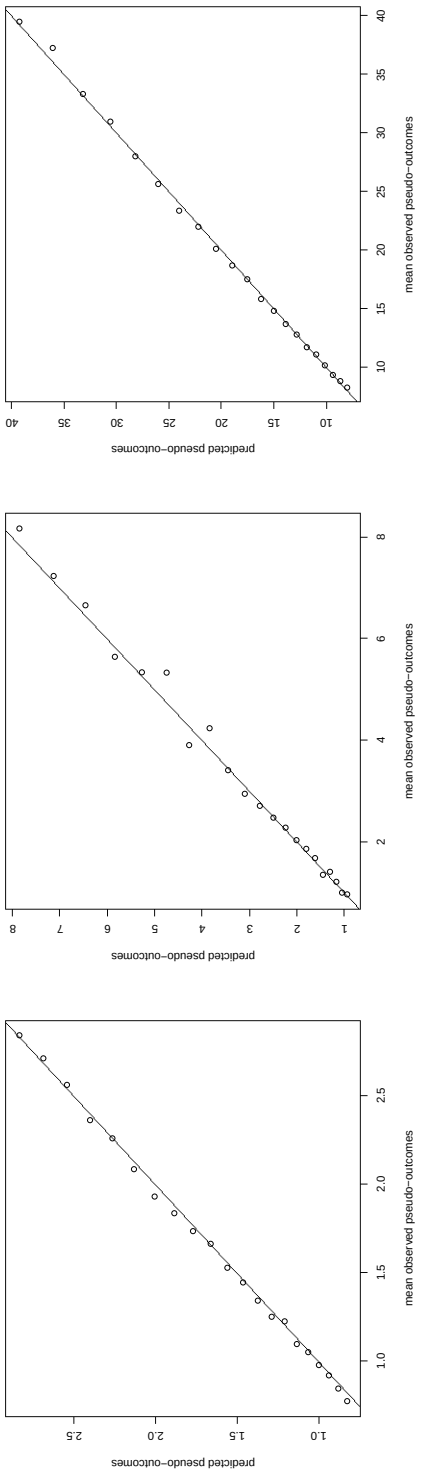


Figure 4.4: Model diagnosis from: setting 1 (left), setting 2 (middle) and setting 3 (right). The x-axis is the mean of the observed pseudo-outcomes from 1000 Monte Carlo runs and the y-axis is the average of the predicted pseudo-outcomes from model (4.23) based on 1000 Monte Carlo runs.

4.3.3 Simulation for clustered data

Recall that the data of the case study are clustered. To study the performance of model (4.23), we now simulate clustered data. In this simulation study, again a univariate predictor X is considered. The response Y is generated from Poisson or negative binomial with

$$\mathrm{E}\left(Y_{ik} \mid X_i, b_i\right) = \exp\left(b_i + \alpha_0 + \alpha_1 X_i\right),$$

where b_i is the sampling location specific random effect. In this simulation study we consider the quadratic model (4.23) and assess: (1) the behaviour of the estimators of the model parameters and (2), whether model approximation (4.23) is appropriate when overdispersions present. In Table 4.10 an overview of the simulation procedure is given.

Table 4.10: Schematic overview of the simulation procedures for clustered data.

1. Generate a series of X_i varying from -1 to 1 in steps of 0.1.
2. Generate a series of b_i from normal distribution with mean 0 and standard deviation 1.
3. Replicate each X_i and b_i for K times.
4. The response $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iK})$ is generated from
   (a) setting 1: Poisson distribution with mean $\lambda(X_i) = \exp(b_i + \alpha_0 + \alpha_1 X_i)$.
   (b) setting 2: negative binomial distribution with mean $\lambda(X_i) = \exp(b_i + \alpha_0 + \alpha_1 X_i)$ and dispersion parameter $\theta = 1$
   (c) setting 3:negative binomial distribution with mean $\lambda(X_i) = \exp(b_i + \alpha_0 + \alpha_1 X_i)$ and dispersion parameter $\theta = 10$
5. Fit model (4.23) to estimate $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$ from estimating equation (4.12) and to estimate the variance-covariance matrix of $\boldsymbol{\beta}$ from the sandwich estimator.

To obtain a consistent estimate of the true β_0 and β_1 we proposed similar approach as before with details given in Table 4.11.

Table 4.11: Schematic overview of obtaining the true regression coefficients β_0 and β_1 for clustered data.

1. Generate a series of n X_i varying from -1 to 1 in steps of 0.1, $n = 21$.
2. Generate a vector b, as same length of X_i, from normal distribution with mean 0 and standard deviation 1 .
3. Replicate each X_i and b_i K times.
4. To obtain β_0, we randomly generate 10^6 Poisson responses Y_{ik} and Y_{il} with mean $\lambda(X_i = 0) = \exp(b_i + \alpha_0)$. We approximate β_0 by replacing the expectation (4.19) in Table 4.2 by its empirical counterpart.
5. To obtain β_1 we randomly generated two 10^6 response vectors Y_{ik} and Y_{il} with mean $\lambda(X_i = x_i) = \exp(b_i + \alpha_0 + \alpha_1 x_i)$ and the corresponding dispersion parameter θ. Generate another two 10^6 vectors of abundance Y_{ik} and Y_{il} with mean $\lambda(X_i = -x_i) = \exp(b_i + \alpha_0 - \alpha_1 x_i)$ and the corresponding dispersion parameter θ. We approximate β_1 by replacing the expectation (4.26) in Table 4.6 by its empirical counterpart.

Simulation result for clustered data

Figure 4.5 shows a linear relationship of $Q(X)$ and X for all three settings. Similar as before, $Q(X)$ is a linear function of X, this indicates that model (4.23) is appropriately specified for modelling the conditional L2 moment. Table 4.12, 4.13 and 4.14 give the numerical result from the simulation studies for setting 1, 2 and 3 respectively. The sandwich variance-covariance estimator corresponds closely to the empirical estimate variance-covariance, and even when data exhibit overdispersion. The sandwich variance-covariance estimator behaves reasonably well. Besides, the sandwich estimator now underestimates the true variance, thus causes more liberal confidence

intervals. Figure 4.6 provides some insights of the model diagnosis, in which the mean observed pseudo-outcomes are plotted against the predicted ones. We see that the points are more sparse around the diagonal line compared to Figure 4.4 for the i.i.d. setting, this implies that clustering in the data may cause less prediction accuracy of model (4.23). However, model (4.23) is an approximation of the L2 moment, model (4.23) still results in feasible predictions.

Table 4.12: Simulation result for setting 1 of the clustered design, based on 1000 Monte Carlo simulation runs. $\mathrm{Av}(\bar{\boldsymbol{\beta}})$ is the average $\bar{\boldsymbol{\beta}} = (\bar{\beta}_0, \bar{\beta}_1)^T$; $\mathrm{Av}(\hat{\boldsymbol{\beta}})$ the average of $\hat{\boldsymbol{\beta}}$; $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ the empirical estimate of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$; $\mathrm{Av}(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$ the estimate of the variance-covariance matrix using sandwich estimator and EC the empirical coverage of a 95% confidence interval for β_0 and β_1.

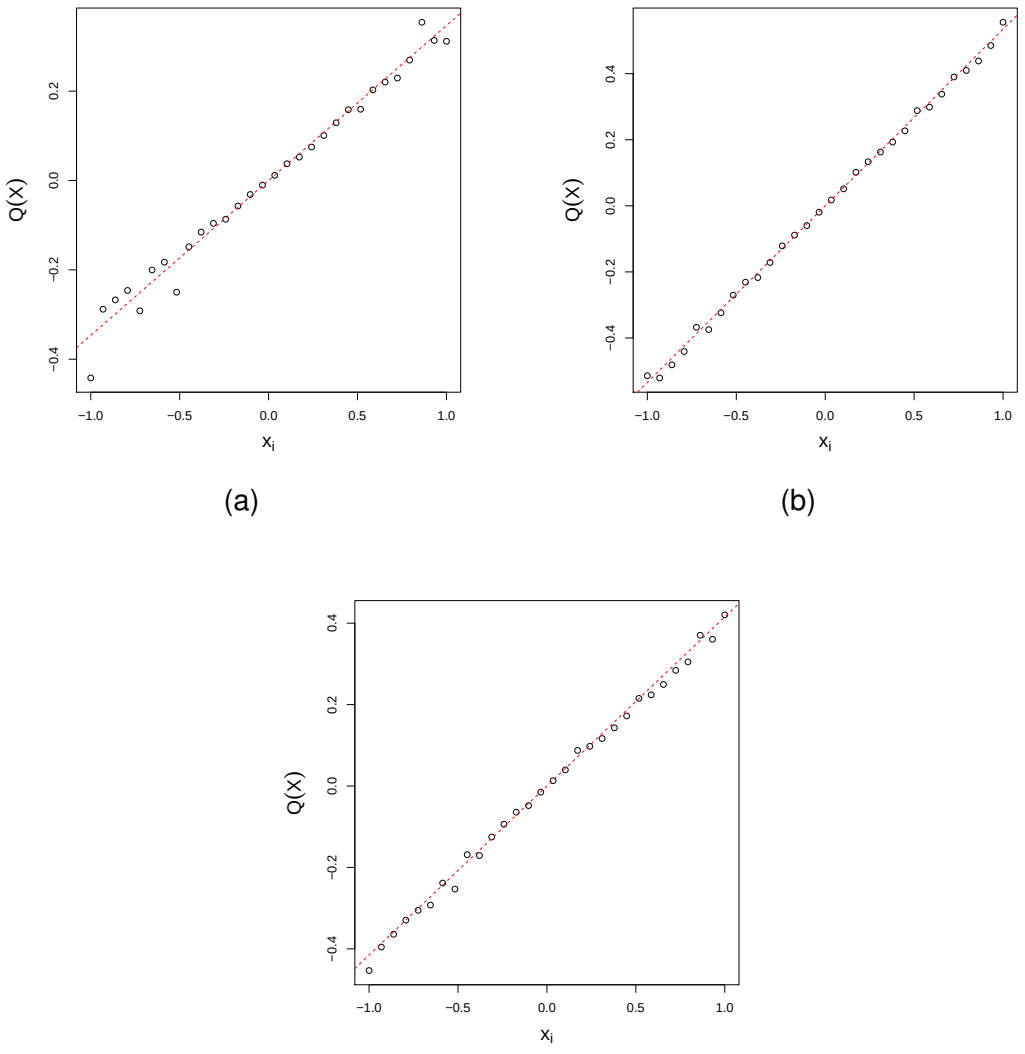
K	$\mathrm{Av}(\bar{\boldsymbol{\beta}})$	$\mathrm{Av}(\hat{\boldsymbol{\beta}})$	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\mathrm{Av}(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$	EC(%)
5	$\begin{bmatrix} 0.67 \\ 0.35 \end{bmatrix}$	$\begin{bmatrix} 0.5503 \\ 0.3197 \\ -0.0337 \end{bmatrix}$	$\begin{bmatrix} 0.0374 & 0.0005 & -0.0315 \\ 0.0005 & 0.0130 & -0.0026 \\ -0.0315 & -0.0026 & 0.0458 \end{bmatrix}$	$\begin{bmatrix} 0.0326 & 0.0007 & -0.0251 \\ 0.0007 & 0.0097 & -0.0030 \\ -0.0251 & -0.0030 & 0.0336 \end{bmatrix}$	$\begin{bmatrix} 85.2 \\ 87.5 \end{bmatrix}$
10	$\begin{bmatrix} 0.67 \\ 0.35 \end{bmatrix}$	$\begin{bmatrix} 0.6501 \\ 0.3234 \\ -0.0193 \end{bmatrix}$	$\begin{bmatrix} 0.0302 & 0.0002 & -0.0230 \\ 0.0002 & 0.0103 & -0.0021 \\ -0.0230 & -0.0021 & 0.0332 \end{bmatrix}$	$\begin{bmatrix} 0.0255 & 0.0005 & -0.0197 \\ 0.0005 & 0.0076 & -0.0022 \\ -0.0197 & -0.0022 & 0.0264 \end{bmatrix}$	$\begin{bmatrix} 91.5 \\ 87.8 \end{bmatrix}$
15	$\begin{bmatrix} 0.67 \\ 0.35 \end{bmatrix}$	$\begin{bmatrix} 0.7021 \\ 0.3224 \\ -0.0348 \end{bmatrix}$	$\begin{bmatrix} 0.0285 & 0.0009 & -0.0223 \\ 0.0007 & 0.0072 & -0.0023 \\ -0.0223 & -0.0018 & 0.0305 \end{bmatrix}$	$\begin{bmatrix} 0.0247 & 0.0007 & -0.0192 \\ 0.0018 & 0.0029 & -0.0034 \\ -0.0192 & -0.0023 & 0.0255 \end{bmatrix}$	$\begin{bmatrix} 90.7 \\ 89.5 \end{bmatrix}$

Table 4.13: Simulation result for setting 2 of the clustered design, based on 1000 Monte Carlo simulation runs. $\mathrm{Av}(\bar{\boldsymbol{\beta}})$ is the average $\bar{\boldsymbol{\beta}} = (\bar{\beta}_0, \bar{\beta}_1)^T$; $\mathrm{Av}(\hat{\boldsymbol{\beta}})$ the average of $\hat{\boldsymbol{\beta}}$; $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ the empirical estimate of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$; $\mathrm{Av}(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$ the estimate of the variance-covariance matrix using sandwich estimator and EC the empirical coverage of a 95% confidence interval for β_0 and β_1.

K	$\mathrm{Av}(\bar{\boldsymbol{\beta}})$	$\mathrm{Av}(\hat{\boldsymbol{\beta}})$	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\mathrm{Av}(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$	EC(%)
5	$\begin{bmatrix} 1.26 \\ 0.53 \end{bmatrix}$	$\begin{bmatrix} 1.4176 \\ 0.5888 \\ -0.0854 \end{bmatrix}$	$\begin{bmatrix} 0.1671 & 0.0169 & -0.1486 \\ 0.0169 & 0.0577 & -0.0289 \\ -0.1486 & -0.0289 & 0.2117 \end{bmatrix}$	$\begin{bmatrix} 0.1082 & 0.0146 & -0.0927 \\ 0.0146 & 0.0497 & -0.0395 \\ -0.0927 & -0.0395 & 0.1378 \end{bmatrix}$	$\begin{bmatrix} 85.6 \\ 90.9 \end{bmatrix}$
10	$\begin{bmatrix} 1.26 \\ 0.53 \end{bmatrix}$	$\begin{bmatrix} 1.5254 \\ 0.5684 \\ -0.0634 \end{bmatrix}$	$\begin{bmatrix} 0.1430 & 0.0078 & -0.1272 \\ 0.0078 & 0.0423 & -0.0127 \\ -0.1272 & -0.0127 & 0.1757 \end{bmatrix}$	$\begin{bmatrix} 0.0980 & 0.0094 & -0.0829 \\ 0.0094 & 0.0312 & -0.0203 \\ -0.0829 & -0.0203 & 0.1077 \end{bmatrix}$	$\begin{bmatrix} 80.6 \\ 89.6 \end{bmatrix}$
15	$\begin{bmatrix} 1.26 \\ 0.53 \end{bmatrix}$	$\begin{bmatrix} 1.5652 \\ 0.5747 \\ -0.0333 \end{bmatrix}$	$\begin{bmatrix} 0.1275 & 0.0114 & -0.1135 \\ 0.0114 & 0.0421 & -0.0183 \\ -0.1135 & -0.0183 & 0.1572 \end{bmatrix}$	$\begin{bmatrix} 0.0895 & 0.0100 & -0.0759 \\ 0.0100 & 0.0328 & -0.0213 \\ -0.0759 & -0.0213 & 0.1038 \end{bmatrix}$	$\begin{bmatrix} 76.8 \\ 89.9 \end{bmatrix}$

Table 4.14: Simulation result for setting 3 of the clustered design, based on 1000 Monte Carlo simulation runs. $\mathrm{Av}(\bar{\boldsymbol{\beta}})$ is the average $\bar{\boldsymbol{\beta}} = (\bar{\beta}_0, \bar{\beta}_1)^T$; $\mathrm{Av}(\hat{\boldsymbol{\beta}})$ the average of $\hat{\boldsymbol{\beta}}$; Cov$(\hat{\boldsymbol{\beta}})$ the empirical estimate of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$; $\mathrm{Av}(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$ the estimate of the variance-covariance matrix using sandwich estimator and EC the empirical coverage of a 95% confidence interval for β_0 and β_1.

K	$\mathrm{Av}(\bar{\boldsymbol{\beta}})$	$\mathrm{Av}(\hat{\boldsymbol{\beta}})$	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\mathrm{Av}(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$	EC(%)
5	$\begin{bmatrix} 0.79 \\ 0.41 \end{bmatrix}$	$\begin{bmatrix} 0.7854 \\ 0.4368 \\ -0.0122 \end{bmatrix}$	$\begin{bmatrix} 0.0869 & 0.0054 & -0.0761 \\ 0.0054 & 0.0247 & -0.0042 \\ -0.0761 & -0.0042 & 0.1033 \end{bmatrix}$	$\begin{bmatrix} 0.0621 & 0.0047 & -0.0508 \\ 0.0047 & 0.0180 & -0.0067 \\ -0.0508 & -0.0067 & 0.0651 \end{bmatrix}$	$\begin{bmatrix} 88.4 \\ 87.9 \end{bmatrix}$
10	$\begin{bmatrix} 0.79 \\ 0.41 \end{bmatrix}$	$\begin{bmatrix} 0.9067 \\ 0.4331 \\ -0.0062 \end{bmatrix}$	$\begin{bmatrix} 0.0719 & 0.0015 & -0.0651 \\ 0.0015 & 0.0202 & -0.0006 \\ -0.0651 & -0.0006 & 0.0899 \end{bmatrix}$	$\begin{bmatrix} 0.0527 & 0.0035 & -0.0431 \\ 0.0035 & 0.0152 & -0.0053 \\ -0.0431 & -0.0053 & 0.0554 \end{bmatrix}$	$\begin{bmatrix} 88.2 \\ 88.2 \end{bmatrix}$
15	$\begin{bmatrix} 0.79 \\ 0.41 \end{bmatrix}$	$\begin{bmatrix} 0.9475 \\ 0.4286 \\ 0.0037 \end{bmatrix}$	$\begin{bmatrix} 0.0752 & 0.0015 & -0.0665 \\ 0.0015 & 0.0186 & 0.0009 \\ -0.0665 & 0.0009 & 0.0882 \end{bmatrix}$	$\begin{bmatrix} 0.0524 & 0.0030 & -0.0428 \\ 0.0030 & 0.0143 & -0.0038 \\ -0.0428 & -0.0038 & 0.0544 \end{bmatrix}$	$\begin{bmatrix} 84.5 \\ 90.1 \end{bmatrix}$



(a) (b)

(c)

Figure 4.5: $Q(X)$, from equation (4.27), is plotted as a function of X for the three simulation settings: (a) for setting 1, (b) for setting 2 and (c) for setting 3 for clustered data. The dashed line is the fitted line from regressing $Q(X)$ against X.

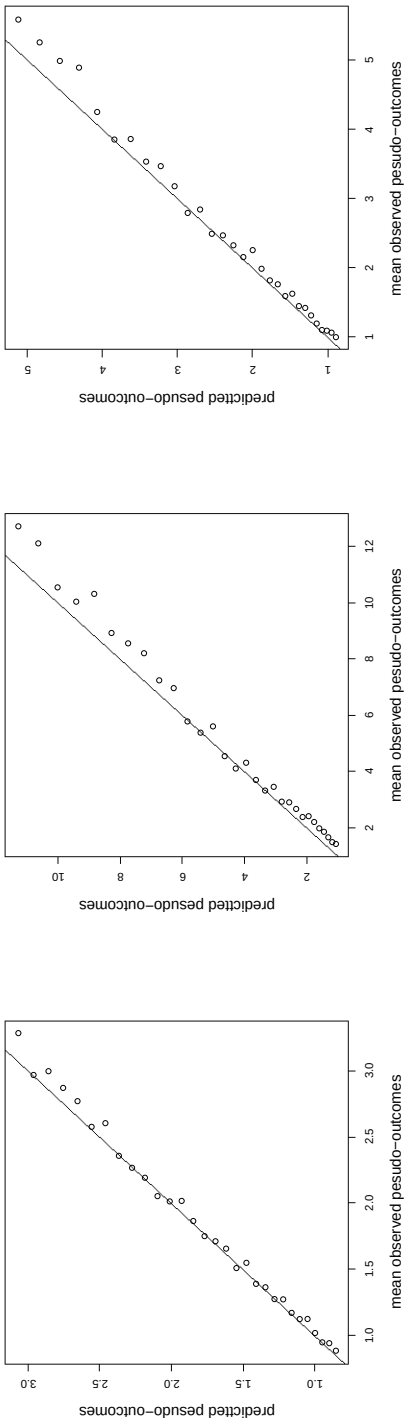


Figure 4.6: Model diagnosis from: setting 1 (left), setting 2 (middle) and setting 3 (right). The x-axis is the mean of the observed pseudo-outcomes from 1000 Monte Carlo runs and the y-axis is the average of the predicted pseudo-outcomes based on model (4.23) from 1000 Monte Carlo runs.

4.4 Example

In this section, we present a case study where we want to model the biodiversity as a function of a continuous variable. The data originate from a community ecology study [**??**]. The study was carried out on the southern shore of a small Laurentide lake on the station de Biologie des Laurentides, Canada. The objective of this application is to learn the reaction of the mite diversity on the water content. The sampling area is a 10×2.6 m transect vegetation mat surrounding the lake and it ranges from the forest border to the water. The sampling area is geographically subdivided into 7 regions by vegetation type. The vegetation types are called 'substratumtype' in the original paper [**?**]. Within each region, several representative spots of 2×2 cm are collected. The resulting data contain the abundance matrix and the environmental data matrix. The oribatid mite abundance matrix contains the counts of 35 adult orbited mites at 70 different spots. The environmental data matrix is composed of the density of the substratum, the water content, the substratum type (Sphagnum magellanicum, Sphagnum rubellum, Sphagnum nemoreum, Sphagnum rubellum $+$ Sphagnum magellanicum, ligneous litter, bare peat and an interface between Sphagnum species) and the coverage density of the shrub. A summary of the data is given in Appendix B.3. **?** have focused on learning the contribution of the substratum type to the mite community variation through canonical correspondence analysis [**?**]. The result showed that only 13.7% of the mite community variation was explained by the substratum variables and the relationship was not significant. Here we will model the L2 moment as a function of a covariate. We will only consider a single covariate since the simulation study was restricted to this setting. Analysing the L2 moment as a function of multiple covariates is beyond the scope of this dissertation. Prior to data analysis, variable water content (X)

is standardized by first subtracting the mean, 405 g/L, and then dividing the standard deviation, which is 134.

4.4.1 Model L1 moment

Consider the following quadratic L1 moment model

$$\mathrm{E}\left(Y_{ik} \mid X_i\right) = \exp\left(\alpha_0 + \alpha_1 X_i + \alpha_2 X_i^2\right), \tag{4.28}$$

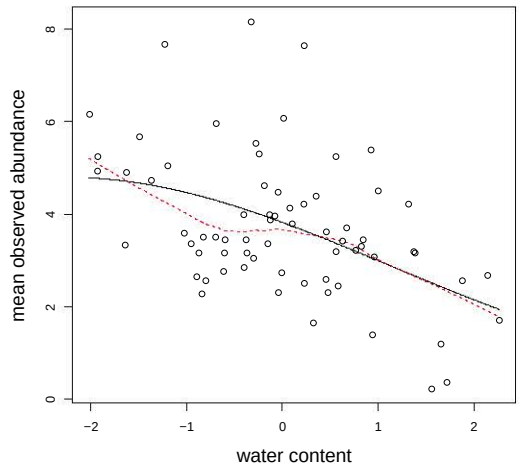
where Y_{ik} denotes the abundance of the oribatid mite type k at location i, $k = 1, \ldots, 35$ and $i = 1, \ldots, 70$. Here X_i denotes the water content at location i. Table 4.15 displays the result of parameter estimation from the fitted model. From the p-value, we see that water content has a significant effect on the mean abundance at 5% level of significance.

Table 4.15: The L1 moment model (4.28) fit.

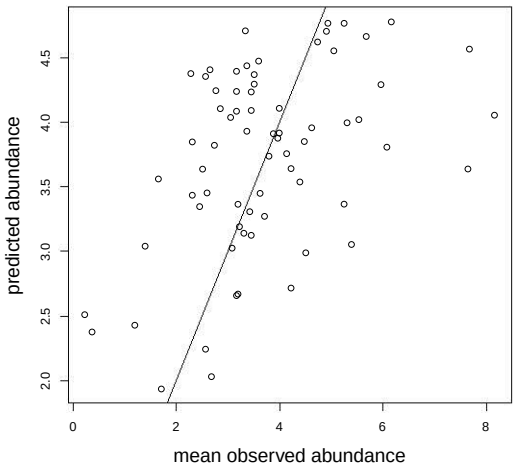
	estimate	standard error	p-value
α_0	1.3403	0.0556	< 0.001
α_1	-0.2002	0.0456	< 0.001
α_2	-0.0439	0.0325	0.1759

Figure 4.7(a) shows the mean observed abundance as a function of the water content. The solid line is the fitted curve while the dashed line is from nonparametric fit using locally-weighted polynomial regression with 0.5 smoother spans. On average we see that the mean abundance is lower at sites where the water content is higher. From Figure 4.7(b) the predicted mean abundance is plotted as a function of the average of the observed abundance at each sampling location. We see that the dots surround the diagonal line, but there is a lot of variability around the line. This is not surprising

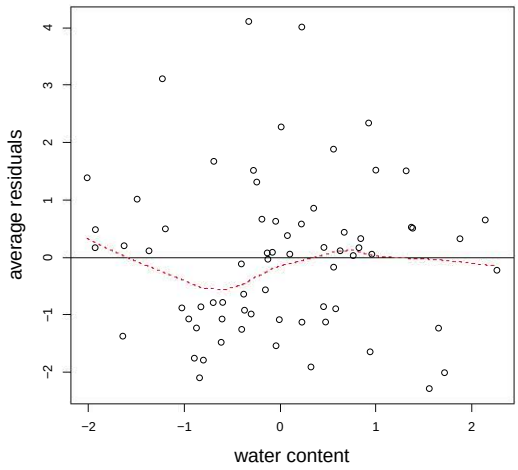
since we only use one explanatory variable. A model diagnostic is carried out and presented in Figure 4.7(c) for which the mean residual is defined as the average of the residuals for each sampling location. The smooth curve shows that neither clear systematic pattern nor serious departure from the horizontal line is detected. So we conclude that model (4.15) properly describe the relationship of the L1 moment and the water content, moreover the water content has a significant effect on the L1 moment.



(a)



(b)



(c)

Figure 4.7: Figure (a): the mean observed abundance as a function of water content. The solid curve is the fitted curve from model (4.28) and the dashed line is from nonparametric fit. Figure (b): the predicted mean abundance is the prediction from L1 moment model (4.28), the mean observed abundance is the average abundance of all the species for each location, the solid line is the diagonal line. Figure (c): the average residuals is defined as the mean of the difference between observed abundance and the predicted abundance. The dashed line is from nonparametric fit.

4.4.2 Model L2 moment

We fit the following L2 moment model to the data:

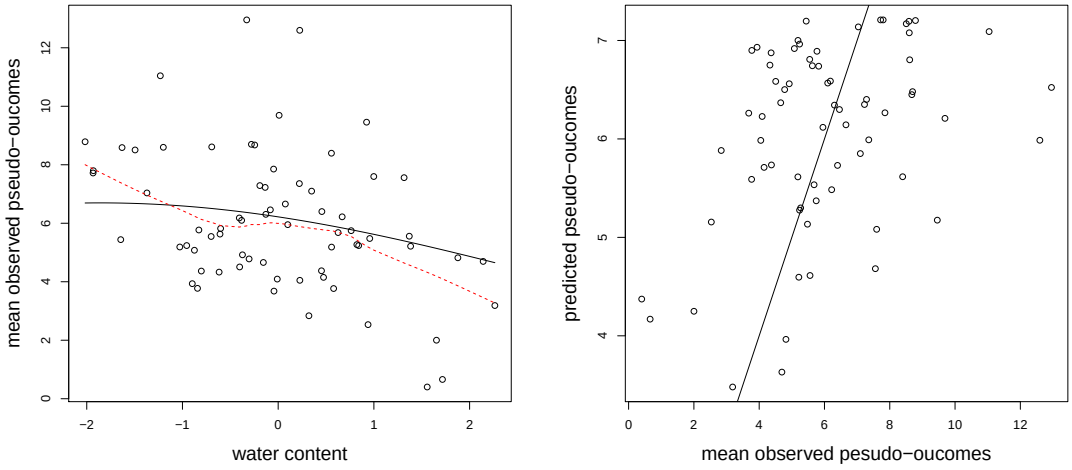
$$\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i\right) = \exp\left[\beta_0 + \beta_1(X_i + X_i) + \beta_2(X_i^2 + X_i^2)\right], \tag{4.29}$$

where $k < l$. Table 4.16 gives the results of the fitted model. From the p-value, it can be seen that water content has a significant effect on the L2 moment at the 5% level of significance.

Table 4.16: The L2 moment model (4.29) fit.

	estimate	standard error	p-value
β_0	1.8276	0.0353	< 0.001
β_1	-0.0797	0.0144	< 0.001
β_2	-0.0214	0.0113	0.1838

In Figure 4.8(a) the mean observed pseudo-outcomes at location i is the average of all the observed pseudo-outcomes related to location i, i.e., the average of $|Y_{ik} - Y_{il}|$. From the fitted line (solid) we see that the mean observed pseudo-outcomes decreases with increase in the water content, and the nonparametric fit (dashed line) resulting from using locally-weighted polynomial regression also shows this negative relation. However, there is still quite a lot of variability around the fitted line. In Figure 4.8(b) the mean observed pseudo-outcomes are plotted against the predicted pseudo-outcomes from model (4.29), the solid line is the diagonal line. We see the points scattered around the diagonal line but, similar as shown in Figure 4.8(a), there still exists a lot of variability. This is due to the fact that we attempt to model the L2 moment using single covariate.



(a) (b)

Figure 4.8: Figure (a): the mean observed pseudo-outcomes as a function of water content. The solid curve is the fitted curve from model (4.29). The dashed line is from nonparametric fit. Figure (b): the predicted pseudo-outcome is the prediction from L2 moment model (4.29), the mean observed pseudo-outcomes are the average pseudo-outcomes related to the location, the solid line is the diagonal line.

4.4.3 Model L1 and L2 moment simultaneously

Recall that the objective is to model the Gini index, which can be derived from the L1 moment model and the L2 moment model as follows:

$$2G_s(X_i) = \frac{\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i, X_i\right)}{\mathrm{E}\left(Y_{ik} \mid X_i\right)}.$$

$G_s(X_i)$ is used to denote the Gini index of site i and the Gini index is a function of the regressor X_i and the superscript 's' stands for 'semiparametric'. As both models are built on the logarithmic link function, this becomes an advantage when modelling the Gini

index. Integrating model (4.28) and (4.29) we have the following model formulation:

$$\begin{aligned} 2G_s(X_i) &= \frac{\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i\right)}{\mathrm{E}\left(Y_{ik} \mid X_i\right)} = \frac{\exp\left(\beta_0 + \beta_1(X_i + X_i) + \beta_2(X_i^2 + X_i^2)\right)}{\exp\left(\alpha_0 + \alpha_1 X_i + \alpha_2 X_i^2\right)} \\ &= \exp\left(\beta_0 - \alpha_0 + (2\beta_1 - \alpha_1)X_i + (2\beta_2 - \alpha_2)X_i^2\right) \\ &= \exp\left(\gamma_0 + \gamma_1 X_i + \gamma_2 X_i^2\right), \end{aligned} \tag{4.30}$$

where $\gamma_0 = \beta_0 - \alpha_0$, $\gamma_1 = 2\beta_1 - \alpha_1$ and $\gamma_2 = 2\beta_2 - \alpha_2$. Given the relation between $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we can estimate the variance of $\boldsymbol{\gamma}$ by the variance-covariance of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Denote $\hat{\boldsymbol{\gamma}}$ the estimate of $\boldsymbol{\gamma}$, the variance of $\hat{\boldsymbol{\gamma}}$ can be obtained under the law of the variance of a sum of two random variables,

$$\mathrm{Var}(\hat{\gamma}_0) = \mathrm{Var}(\hat{\beta}_0 - \hat{\alpha}_0) = \mathrm{Var}(\hat{\beta}_0) + \mathrm{Var}(\hat{\alpha}_0) - 2\mathrm{Cov}(\hat{\beta}_0, \hat{\alpha}_0).$$

Analogously we can estimate the variance of $\hat{\gamma}_1$ in terms of variance-covariance of $\hat{\alpha}_1$ and $\hat{\beta}_1$ as

$$\mathrm{Var}(\hat{\gamma}_1) = \mathrm{Var}(2\hat{\beta}_1 - \hat{\alpha}_1) = 4\mathrm{Var}(\hat{\beta}_1) + \mathrm{Var}(\hat{\alpha}_1) - 4\mathrm{Cov}(\hat{\beta}_1, \hat{\alpha}_1).$$

Similarly,

$$\mathrm{Var}(\hat{\gamma}_2) = \mathrm{Var}(2\hat{\beta}_2 - \hat{\alpha}_2) = 4\mathrm{Var}(\hat{\beta}_2) + \mathrm{Var}(\hat{\alpha}_2) - 4\mathrm{Cov}(\hat{\beta}_2, \hat{\alpha}_2).$$

To be able to further study the statistical properties of the $\hat{\gamma}$ estimators, we will need to find estimator of the covariance between $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$. In next section, we introduce the concept of influence function which makes it possible to estimate this covariance.

4.5 Inference of the Gini index model

This section contains the estimation of the covariance of the L1 and L2 moment model parameter estimators. A key element of the derivation of this covariance estimator is the use of influence functions. In Section 4.5.1, an introduction of influence function is given. We derive the influence function for the L1 moment model in Section 4.5.2. In Section 4.5.3 we work out the expression of the influence function for the L2 moment model. Finally in Section 4.5.4 an estimator of the covariance is obtained based on influence functions. The inference of the influence function based variance estimator studied in this section is only for i.i.d. setting, since influence function based variance estimator for the clustered setting falls beyond the scope of this dissertation.

4.5.1 Introduction

We give a general introduction to the influence function of an asymptotically linear estimator in this section, we refer the readers to Chapter 3 of **?** and Chapter 7 of **?** for more details.

Consider i.i.d. random vectors $\{\boldsymbol{Z}_i = (Y_i, \boldsymbol{X}_i) \mid i = 1, \ldots, n\}$ and then denote the density of a single $\boldsymbol{Z}$ by $f_{\boldsymbol{Z}}(\boldsymbol{z}; \boldsymbol{\beta})$. Let $\boldsymbol{\beta}_0$ denote the true $\boldsymbol{\beta}$, which is q-dimensional, and $\hat{\boldsymbol{\beta}}$ denote an estimator of $\boldsymbol{\beta}$.

Definition 2 (Influence function)**.** $\hat{\boldsymbol{\beta}}$ *is* asymptotically linear *if there exists a* q*-dimensional measurable random function* $\boldsymbol{\varphi}(\boldsymbol{Z})$ *and the* i^{th} *element of* $\boldsymbol{\varphi}$ *measures the influence of the* i^{th} *observation on the estimator* $\hat{\boldsymbol{\beta}}$*. Such a random vector is of mean zero, i.e.,*

$\mathrm{E}[\boldsymbol{\varphi}(\boldsymbol{Z})] = \mathbf{0}$*, and*

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{\varphi}(\boldsymbol{Z}_i) + o_p(1), \tag{4.31}$$

where $o_p(1)$ *is a term that converges in probability to zero as* n *goes to infinity and* $\mathrm{E}\left(\boldsymbol{\varphi}\boldsymbol{\varphi}^T\right)$ *is finite and nonsingular. The random function* $\boldsymbol{\varphi}(\cdot)$ *is called influence function.*

As the sample size goes to infinity, most of the reasonable estimators of β are asymptotically linear estimators and such an estimator possesses an unique influence function (Theorem 3.1, **?**). The following asymptotic property of an asymptotically linear estimator $\hat{\beta}$ is also give by **?**:

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N\left[\mathbf{0}, \mathrm{E}\left(\boldsymbol{\varphi}\boldsymbol{\varphi}^T\right)\right].$$

Therefore the variance of $\hat{\beta}$ can be estimated via its influence function. In the next sections the influence functions of parameter estimators of the L1 moment model and the L2 moment model are given. The influence functions give rise to the estimator of the variance-covariance matrix of the parameter estimators of the Gini index model.

4.5.2 Influence function for L1 moment

As introduced in Section 4.2.2, when the response variable are species abundance, a general form of the L1 moment model is

$$\mathrm{E}\left(Y_{ik} \mid \boldsymbol{X}_i\right) = g^{-1}(\boldsymbol{X}_i^T\boldsymbol{\alpha}) = \exp\left(\boldsymbol{X}_i^T\boldsymbol{\alpha}\right), \tag{4.32}$$

where $i = 1, \ldots, n$, $k = 1, \ldots, K$ and $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$. Following Equation (4.9) in **?**, an arbitrary influence function for the L1 moment model (4.32) is given as follows:

$$\boldsymbol{\varphi}_{l1} = \Big[\mathrm{E}\left(\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X}\right)\Big]^{-1}\boldsymbol{X}_i^T\Big[\boldsymbol{Y}_i - \exp(\boldsymbol{X}_i\boldsymbol{\alpha}_0)\Big], \tag{4.33}$$

$\boldsymbol{\alpha}_0$ is the truth, $\boldsymbol{R}$ is a diagonal matrix with elements $\exp(\boldsymbol{X}\boldsymbol{\alpha}_0)$ and $\mathrm{E}(\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X})$ can be estimated empirically. The estimator $\hat{\boldsymbol{\alpha}}$ possesses the following asymptotic property [**?**]:

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \xrightarrow{D} N(\mathbf{0}, \mathrm{E}(\boldsymbol{\varphi}_{l1}\boldsymbol{\varphi}_{l1}^T)).$$

The expression of the $\mathrm{Var}(\hat{\boldsymbol{\alpha}})$ in terms of the influence function is given in Appendix A.3.3.

4.5.3 Influence function for L2 moment model

Let $(Y, \boldsymbol{X})$ and $(Y,' \boldsymbol{X}')$ denote i.i.d. random observations, consider the following L2 moment model

$$\mathrm{E}\left(|Y - Y'| \mid \boldsymbol{X}, \boldsymbol{X}'\right) = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}) = \exp\Big[(\boldsymbol{X} + \boldsymbol{X}')^T\boldsymbol{\beta}\Big] \quad (\boldsymbol{X}, \boldsymbol{X}') \in \mathcal{X}_0, \tag{4.34}$$

where $\mathcal{X}_0$ denote the covariate space where no restriction is imposed. Denote $\boldsymbol{u}_2$ the estimating equation for model (4.34). Following the form of estimating equation (4.12) of the restricted moment model in Section 4.2 we have

$$\boldsymbol{u}_2(\boldsymbol{X}, Y, \boldsymbol{X}', Y') = \sum_{i=1}^{n}\sum_{h=1}^{n}\boldsymbol{B}(\boldsymbol{X}_i, \boldsymbol{X}_h; \boldsymbol{\beta})\Big[\,|Y_i - Y_h| - m(\boldsymbol{X}_i, \boldsymbol{X}_h; \boldsymbol{\beta})\Big], \tag{4.35}$$

where

$$\boldsymbol{B}(\boldsymbol{X}_i, \boldsymbol{X}_h; \boldsymbol{\beta}) = \frac{\partial m(\boldsymbol{X}_i, \boldsymbol{X}_h; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}.$$

Following Theorem 2 of **?**, the influence function of the parameter estimators in model (4.34), denoted by $\boldsymbol{\varphi}_{l2}$, can be expressed as:

$$\boldsymbol{\varphi_{l2}}(\boldsymbol{X}_i, Y_i) = \tag{4.36}$$

$$\boldsymbol{C}\mathrm{E}\left\{(\boldsymbol{X}_i + \boldsymbol{X}_h)\exp\left[(\boldsymbol{X}_i + \boldsymbol{X}_h)^T\boldsymbol{\beta}_0\right]\left\{|Y_i - Y_h| - \exp\left[(\boldsymbol{X}_i + \boldsymbol{X}_h)^T\boldsymbol{\beta}_0\right]\right\} \mid \boldsymbol{X}_i, Y_i\right\},$$

where the conditional expectation can be replaced by a sample average over index i and

$$\begin{aligned}\boldsymbol{C} &= \mathrm{E}\Big\{(\boldsymbol{X}_i + \boldsymbol{X}_h)\exp^2\left[(\boldsymbol{X}_i + \boldsymbol{X}_h)\boldsymbol{\beta}_0\right] \\ &\quad \left\{|Y_i - Y_h| - \exp\left[(\boldsymbol{X}_i + \boldsymbol{X}_h)^T\boldsymbol{\beta}_0\right]\right\}^2 (\boldsymbol{X}_i + \boldsymbol{X}_h)^T \mid \boldsymbol{X}_i, \boldsymbol{X}_h, Y_i, Y_h\Big\}^{-1},\end{aligned}$$

and $\boldsymbol{C}$ can be estimated empirically over indices i and h. We again refer to Section 7.3.4 of **?** for the proof of the following asymptotic properties of estimator $\hat{\beta}$,

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}),$$

and

$$\tilde{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = \boldsymbol{C}\mathrm{E}\left[\boldsymbol{K}(\boldsymbol{X}_i, Y_i, \boldsymbol{\beta}_0)\boldsymbol{K}^T(\boldsymbol{X}_i, Y_i, \boldsymbol{\beta}_0)\right]\boldsymbol{C}^T$$

with

$$\begin{aligned}\boldsymbol{K}(\boldsymbol{X}_i, Y_i, \boldsymbol{\beta}_0) &= \mathrm{E}\Big\{(\boldsymbol{X}_i + \boldsymbol{X}_h)\exp[(\boldsymbol{X}_i + \boldsymbol{X}_h)^T\boldsymbol{\beta}_0] \\ &\quad \Big\{|Y_i - Y_h| - \exp\Big[(\boldsymbol{X}_i + \boldsymbol{X}_h)^T\boldsymbol{\beta}_0\Big]\Big\} \mid \boldsymbol{X}_i, Y_i\Big\}.\end{aligned}$$

The expression of the $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ in terms of the influence function is given in Appendix A.3.3.

4.5.4 An estimate of the covariance

Let $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ denote the estimator of the L1 and L2 moment model respectively; see section 4.5.2 and 4.5.3. As discussed in Section 4.5.1 we can write these estimators as a function of their influence functions.

$$\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\varphi}_{l1}(Y_i, \boldsymbol{X}_i) + o_p(1),$$

and

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\varphi}_{l2}(Y_i, \boldsymbol{X}_i) + o_p(1).$$

From the expression it follows that:

$$\begin{aligned}n\mathrm{Cov}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) &= n\mathrm{Cov}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \qquad (4.37)\\ &= \mathrm{Cov}[\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)] \\ &= \mathrm{Cov}[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\varphi}_{l1}(Y_i, \boldsymbol{X}_i), \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\varphi}_{l2}(Y_i, \boldsymbol{X}_i)] + o_p(1) \\ &= \mathrm{Cov}[\boldsymbol{\varphi}_{l1}(Y_i, \boldsymbol{X}_i), \boldsymbol{\varphi}_{l2}(Y_i, \boldsymbol{X}_i)].\end{aligned}$$

It therefore follows that $n\mathrm{Cov}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ ca be estimated by:

$$\frac{1}{n-1}\sum_{i=1}^{n}[(\boldsymbol{\varphi}_{l1}(Y_i, \boldsymbol{X}_i) - \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{\varphi}_{l1}(Y_j, \boldsymbol{X}_j)][\boldsymbol{\varphi}_{l2}(Y_i, \boldsymbol{X}_i) - \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{\varphi}_{l2}(Y_j, \boldsymbol{X}_j)].$$

Since we now have an estimator of $\mathrm{Cov}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$, $\mathrm{Var}(\hat{\boldsymbol{\gamma}})$ of the Gini index model can be then estimated; see Section 4.4.3.

4.6 Simulation study

Recall the following L1 moment model we proposed in Section 4.3:

$$\mathrm{E}\left(Y_i \mid X_i\right) = \exp(\alpha_0 + \alpha_1 X_i),$$

and the L2 moment model:

$$\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i\right) = g^{-1}\left[\beta_0 + \beta_1\left(X_i + X_i\right) + \beta_1\left(X_i^2 + X_i^2\right)\right]. \tag{4.38}$$

where $k < l$. A specific property of this model is that the pseudo-outcomes were restricted to within sampling location comparisons, we use $\mathcal{X}_c$ to denote the covariate set for model (4.38). The theory from Section 4.5, however, is valid for the L2 moment model that is defined for all possible pairwise comparisons. Equivalently, model (4.38) can be rewritten as:

$$\mathrm{E}(|Y_{ik} - Y_{hl}| \mid X, X') = \exp\left\{\beta_{00}\,(1 - I_{ih}) + \left[\beta_0 + \beta_1(X_{ik} + X_{hl}) + \beta_2(X_{ik}^2 + X_{hl}^2)\right] I_{ih}\right\}, \tag{4.39}$$

where $(\boldsymbol{X}, \boldsymbol{X}') \in \mathcal{X}_0$ and $I_{ih} = 1$ if $i = h$ and 0 otherwise. Model (4.39) actually has two parts:

$$\mathrm{E}(|Y_{ik} - Y_{hl}| \mid X_i, X_h) = \begin{cases} \exp(\beta_{00}) & \text{: for observations from different locations.} \\ \exp\left[\beta_0 + \beta_1(X_{ik} + X_{hl}) + \beta_2(X_{ik}^2 + X_{hl}^2)\right] & \text{: for observations from the same location.} \end{cases}$$

We consider β_{00} as a nuisance parameter, since it is not directly of interest. It is a consequence of writing model (4.39) as a function of $\mathcal{X}_0$ instead of $\mathcal{X}_c$.

4.6.1 Simulation for i.i.d. setting

We assess the parameter estimation of model (4.39) in a simulation study where we consider i.i.d. setting. The simulation scheme is given in Table 4.17

In this simulation study we use the same data generation procedures as in Section 4.3, but with a focus on setting 1 where the data are generated from Poisson distribution. The goal is to assess the behaviour of the influence function based variance estimators proposed in Section 4.5.3. A simulation study with data generated from setting 1 is sufficient for the purpose of assessing the behaviour of the parameter estimators. From Table 4.18 we can see that the empirical coverage of a 95% confidence interval for β_0

Table 4.17: Schematic overview of the simulation study for the i.i.d. setting

1. Generate a series of equispaced X_i varying from -1 to 1.
2. Replicate each X_i K times.
3. The response Y_{ik}, $k = 1, \ldots, K$, is generated from Poisson distribution with mean $\lambda(X_i) = \exp(\alpha_0 + \alpha_1 X_i)$
4. Fit model (4.23) using the generated X_i and Y_{ik} to obtain the estimate of $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ from sandwich variance estimator (4.15).
5. Fit model (4.39) using the generated X_i and Y_{ik} to obtain estimate of $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ from the influence based variance estimator.

and β_1 improves with increasing sample size. The results further indicate that the bias in $\hat{\boldsymbol{\beta}}$ decreases with increasing sample size. the influence function based variance estimator corresponds well to the one from using sparse correlation theory (4.15).

Table 4.18: Result of the simulation study for i.i.d. setting based on 1000 Monte Carlo runs. K is the number of replicates; Av($\bar{\boldsymbol{\beta}}$) the average of the unbiased estimate of the true β_0 and β_1; $\mathrm{Av}(\hat{\boldsymbol{\beta}})$ the average of the $\boldsymbol{\beta}$ estimates; $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ the empirical variance of $\hat{\boldsymbol{\beta}}$; $\mathrm{Av}(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$ the average of the sandwich variance estimator (4.15); $\mathrm{Av}(\tilde{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$ the average of the influence function based variance estimator; EC1 the empirical coverage of a 95% confidence interval for β_0 and β_1 based on $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$ and EC2 the coverage calculated based on $\tilde{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$.

K	$\bar{\boldsymbol{\beta}}$	$\mathrm{Av}(\hat{\boldsymbol{\beta}})$	$\mathrm{Var}(\hat{\boldsymbol{\beta}})$	$\mathrm{Av}(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$	$\mathrm{Av}(\tilde{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$	EC1(%)	EC2(%)
5	$\begin{bmatrix} 0.6410 \\ 0.3191 \end{bmatrix}$	$\begin{bmatrix} 0.6124 \\ 0.3275 \\ -0.0295 \end{bmatrix}$	$\begin{bmatrix} 0.0884 & -0.0014 & -0.0522 \\ -0.0014 & 0.0216 & -0.0034 \\ -0.0522 & -0.0034 & 0.0566 \end{bmatrix}$	$\begin{bmatrix} 0.0679 & 0.0005 & -0.0380 \\ 0.0005 & 0.0121 & -0.0053 \\ -0.0380 & -0.0053 & 0.0345 \end{bmatrix}$	$\begin{bmatrix} 0.0822 & 0.0006 & -0.0464 \\ 0.0006 & 0.0152 & -0.0063 \\ -0.0464 & -0.0063 & 0.0431 \end{bmatrix}$	$\begin{bmatrix} 86.1 \\ 90.0 \end{bmatrix}$	$\begin{bmatrix} 84.2 \\ 88.9 \end{bmatrix}$
10	$\begin{bmatrix} 0.6410 \\ 0.3191 \end{bmatrix}$	$\begin{bmatrix} 0.6303 \\ 0.3214 \\ -0.0194 \end{bmatrix}$	$\begin{bmatrix} 0.0320 & -0.0007 & -0.0182 \\ -0.0007 & 0.0077 & -0.0012 \\ -0.0182 & -0.0012 & 0.0201 \end{bmatrix}$	$\begin{bmatrix} 0.0307 & 0.0002 & -0.0180 \\ 0.0002 & 0.0060 & -0.0018 \\ -0.0180 & -0.0018 & 0.0176 \end{bmatrix}$	$\begin{bmatrix} 0.0342 & 0.0002 & -0.0201 \\ 0.0002 & 0.0068 & -0.0019 \\ -0.0201 & -0.0019 & 0.0197 \end{bmatrix}$	$\begin{bmatrix} 92.7 \\ 93.6 \end{bmatrix}$	$\begin{bmatrix} 90.4 \\ 92.3 \end{bmatrix}$
20	$\begin{bmatrix} 0.6410 \\ 0.3191 \end{bmatrix}$	$\begin{bmatrix} 0.6374 \\ 0.3192 \\ -0.0114 \end{bmatrix}$	$\begin{bmatrix} 0.0161 & -0.0002 & -0.0099 \\ -0.0002 & 0.0035 & -0.0006 \\ -0.0099 & -0.0006 & 0.0104 \end{bmatrix}$	$\begin{bmatrix} 0.0147 & 0.0001 & -0.0088 \\ 0.0001 & 0.0030 & -0.0006 \\ -0.0088 & -0.0006 & 0.0088 \end{bmatrix}$	$\begin{bmatrix} 0.0155 & 0.0001 & -0.0093 \\ 0.0001 & 0.0032 & -0.0007 \\ -0.0093 & -0.0007 & 0.0093 \end{bmatrix}$	$\begin{bmatrix} 92.5 \\ 93.1 \end{bmatrix}$	$\begin{bmatrix} 91.8 \\ 92.7 \end{bmatrix}$

4.6.2 Simulation for clustered setting

Since the influence function (4.36) we derived for the L2 moment model is based on i.i.d. random observations, in this section, we assess the behaviour of the influence function based variance estimator under different degrees of clustering in the data. An overview of the simulation procedures is given in Table 4.19.

Table 4.19: Schematic overview of the simulation study

1. Generate a series of equispaced X_i varying from -1 to 1.
2. Generate the random effect b_i from Normal distribution with $b \sim N(0, \sigma)$.
3. Replicate each X_i and b_i K times.
4. The response Y_{ik}, $k = 1, \ldots, K$, is generated from Poisson distribution with mean $\lambda(X_i, b_i) = \exp(\alpha_0 + b_i + \alpha_1 X_i)$
5. Fit model (4.28) using the generated X_i and Y_{ij} to obtain $\hat{\boldsymbol{\alpha}}$.
6. Fit model (4.39) using the generated X_i and Y_{ij} to obtain $\hat{\boldsymbol{\beta}}$.
7. Obtain $\hat{\boldsymbol{\gamma}}$ from its relation with $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ as shown in (4.31) on page 147.
8. Obtain the estimate of the variance of $\hat{\boldsymbol{\gamma}}$ using (4.37).

In this simulation study we again consider the response generated from Poisson distribution. We choose the random effect parameter σ to vary for 5 different levels: $0.5s$, s, $2s$, $4s$ and $8s$ where $s = 0.493$. The choice of s is based on the case study in Section 4.4 and is obtained from fitting the orbited mites data by the following random intercept model:

$$\mathrm{E}\left(Y_{ik} \mid X_i, b_i\right) = \log(\alpha_0 + b_i + \alpha_1 X_i),$$

where X_i is the water content, b_i is a location-specific random effect and the estimated

standard error of b_i is 0.493 from using package 'lme4' [**?**] in R.

We present the empirical estimate of the $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ and $\mathrm{Var}(\hat{\boldsymbol{\gamma}})$, to which we compare the sandwich estimator and the influence function based estimator. Note that we only reported the estimates of the variance of $\hat{\gamma}$, but not the covariance since it is not of direct interest to us in the application study. The estimates with more than 50% bias are given in bold. We conclude that in general when the variance of the random intercept increases, the influence function based variance-covariance estimator (4.37) we derived in Section 4.5.4 for the L2 moment model tends to be biased. However, we see that from Table 4.20, the influence function based variance-covariance estimator corresponds well to the empirical variance-covariance estimator when $\sigma = s/2$ or s.

Table 4.20: Simulation result from 1000 Monte Carlo runs. σ denotes the variance of the random effect and $s = 0.493$; $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ denotes the empirical variance of $\hat{\boldsymbol{\beta}}$ from the L2 moment model; $\mathrm{Av}(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$ denotes the mean estimate of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ using sandwich estimator; $\mathrm{Av}(\tilde{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$ denotes the mean estimate of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ using influence functions; $\mathrm{Var}(\hat{\boldsymbol{\gamma}})$ denotes the empirical variance of $\hat{\boldsymbol{\gamma}}$ and $\mathrm{Av}(\tilde{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\gamma}}})$ is the mean estimate of the variance of $\hat{\boldsymbol{\gamma}}$ based on influence functions. Estimates with more than 50% bias compared to the empirically estimated ones are indicated in bold.

σ	$\mathrm{Var}(\hat{\boldsymbol{\beta}})$	$\mathrm{Av}(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$	$\mathrm{Av}(\tilde{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})$	$\mathrm{Var}(\hat{\boldsymbol{\gamma}})$	$\mathrm{Av}(\tilde{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\gamma}}})$
$\frac{s}{2}$	$\begin{bmatrix} 0.0196 & 0.0002 & -0.0143 \\ 0.0002 & 0.0056 & -0.0011 \\ -0.0143 & -0.0011 & 0.0185 \end{bmatrix}$	$\begin{bmatrix} 0.0147 & \mathbf{0.0005} & -0.0092 \\ \mathbf{0.0005} & 0.0034 & \mathbf{-0.0017} \\ -0.0092 & \mathbf{-0.0017} & 0.0096 \end{bmatrix}$	$\begin{bmatrix} 0.0134 & 0.0001 & -0.0087 \\ 0.0000 & 0.0050 & -0.0015 \\ -0.0087 & -0.0015 & 0.0141 \end{bmatrix}$	$\begin{bmatrix} 0.0215 \\ 0.0206 \\ 0.0814 \end{bmatrix}$	$\begin{bmatrix} 0.0196 \\ 0.0246 \\ 0.0601 \end{bmatrix}$
s	$\begin{bmatrix} 0.0331 & 0.0001 & -0.0234 \\ 0.0001 & 0.0093 & -0.0021 \\ -0.0234 & -0.0021 & 0.0298 \end{bmatrix}$	$\begin{bmatrix} 0.0239 & \mathbf{0.0009} & -0.0151 \\ \mathbf{0.0009} & 0.0056 & -0.0028 \\ -0.0151 & -0.0028 & 0.0159 \end{bmatrix}$	$\begin{bmatrix} \mathbf{0.0163} & 0.0001 & \mathbf{-0.0104} \\ 0.0001 & 0.0060 & -0.0021 \\ \mathbf{-0.0104} & -0.0021 & 0.0165 \end{bmatrix}$	$\begin{bmatrix} 0.0386 \\ 0.0413 \\ 0.1656 \end{bmatrix}$	$\begin{bmatrix} 0.0542 \\ \mathbf{0.06618} \\ 0.1655 \end{bmatrix}$
$2s$	$\begin{bmatrix} 0.0895 & 0.0019 & -0.0641 \\ 0.0019 & 0.0254 & -0.0074 \\ -0.0641 & -0.0074 & 0.0822 \end{bmatrix}$	$\begin{bmatrix} 0.0583 & 0.0022 & -0.0372 \\ 0.0022 & 0.0143 & -0.0076 \\ -0.0372 & -0.0076 & 0.0399 \end{bmatrix}$	$\begin{bmatrix} \mathbf{0.0275} & \mathbf{0.0001} & \mathbf{-0.0170} \\ \mathbf{0.0001} & \mathbf{0.0099} & -0.0042 \\ \mathbf{-0.0170} & -0.0042 & \mathbf{0.0255} \end{bmatrix}$	$\begin{bmatrix} 0.1384 \\ 0.4058 \\ 0.7813 \end{bmatrix}$	$\begin{bmatrix} 0.1783 \\ 0.4648 \\ 0.7771 \end{bmatrix}$
$4s$	$\begin{bmatrix} 0.3598 & 0.0123 & -0.2700 \\ 0.0123 & 0.1364 & -0.0659 \\ -0.2700 & -0.0659 & 0.3799 \end{bmatrix}$	$\begin{bmatrix} \mathbf{0.1775} & \mathbf{0.0032} & \mathbf{-0.1163} \\ \mathbf{0.0032} & 0.0722 & -0.0507 \\ \mathbf{-0.1163} & -0.0507 & \mathbf{0.1688} \end{bmatrix}$	$\begin{bmatrix} \mathbf{0.0679} & \mathbf{-0.0022} & \mathbf{-0.0405} \\ \mathbf{-0.0022} & \mathbf{0.0273} & \mathbf{-0.0147} \\ \mathbf{-0.0405} & \mathbf{-0.0147} & \mathbf{0.0708} \end{bmatrix}$	$\begin{bmatrix} 2.6365 \\ 30.4268 \\ 28.0248 \end{bmatrix}$	$\begin{bmatrix} 1.7006 \\ \mathbf{13.6182} \\ \mathbf{12.2714} \end{bmatrix}$
$8s$	$\begin{bmatrix} 0.3549 & 0.0098 & -0.2618 \\ 0.0098 & 0.1341 & -0.0599 \\ -0.2618 & -0.0599 & 0.3579 \end{bmatrix}$	$\begin{bmatrix} \mathbf{0.1755} & \mathbf{0.0024} & \mathbf{-0.1144} \\ \mathbf{0.0024} & 0.0750 & -0.0513 \\ \mathbf{-0.1144} & -0.0513 & \mathbf{0.1670} \end{bmatrix}$	$\begin{bmatrix} \mathbf{0.0663} & \mathbf{-0.0016} & \mathbf{-0.0404} \\ \mathbf{-0.0016} & \mathbf{0.0263} & \mathbf{-0.0132} \\ \mathbf{-0.0404} & \mathbf{-0.0132} & \mathbf{0.0676} \end{bmatrix}$	$\begin{bmatrix} 2.5660 \\ 29.9315 \\ 26.4592 \end{bmatrix}$	$\begin{bmatrix} 1.4495 \\ \mathbf{12.0134} \\ \mathbf{11.6451} \end{bmatrix}$

4.7 Revisit the example

The influence functions we proposed in Section 4.5 enabled the estimation of the variance of $\hat{\gamma}$ in the Gini index model (4.31). Although the influence function for the L2 moment model is only derived for i.i.d. data, we have assessed that the influence function based estimator (4.37) is a feasible estimator when the correlation within a cluster is small. Since the variance estimator (4.37) can be biased we use a bootstrap procedure to estimate the variance of the model parameter estimators. After 1000 bootstrapping runs the the empirical variances of $\hat{\gamma}$ are reported.

In Table 4.21 the detailed results are shown. The p-values from Wald-type test based on two different variance estimators agree with each other regarding $\hat{\gamma_0}$ and $\hat{\gamma_2}$, but not for $\hat{\gamma_1}$. The model parameters can be explained as follows: when the water content of the substrate is held at its mean, i.e., 405 g/L, the estimate of Gini index is $\frac{\exp(\hat{\gamma}_0)}{2} = 0.8139$. An 1 g/L increase of the water content results in an increase of 1/134=0.0075 g/L increase in the standardized variable, therefore 1 g/L increase of water content result in an increase of the Gini index by $\exp(0.0075\hat{\gamma}_1 + 0.0075^2\hat{\gamma}_2)/2 = 0.5$.

Table 4.21: The model fit of Gini index Model (4.31). $\hat{\boldsymbol{\gamma}}_b$ is the estimated $\boldsymbol{\gamma}$ from bootstrapping; $\hat{\mathrm{se}}(\hat{\boldsymbol{\gamma}})$ the estimated standard error of $\hat{\boldsymbol{\gamma}}$ based on influence functions; $\hat{\mathrm{se}}(\hat{\boldsymbol{\gamma}}_b)$ the empirical standard error of $\hat{\boldsymbol{\gamma}}_b$; p-value the p-value based on $\hat{\mathrm{se}}(\hat{\boldsymbol{\gamma}})$ and p_b-value the p-value based on $\hat{\mathrm{se}}(\hat{\boldsymbol{\gamma}}_b)$.

	$\hat{\boldsymbol{\gamma}}$	$\hat{\mathrm{se}}(\hat{\boldsymbol{\gamma}})$	$\hat{\mathrm{se}}(\hat{\boldsymbol{\gamma}}_b)$	p-value	p_b-value
γ_0	0.4873	0.0232	0.0496	< 0.01	< 0.01
γ_1	0.0409	0.0190	0.0374	0.0310	0.2671
γ_2	0.0011	0.0140	0.0178	0.9377	0.9507

The smooth curve in Figure 4.9(a) shows a positive relation of water content and the $\hat{G}$, i.e., the nonparametric Gini index estimate (4.1). The proposed model fit (the solid line) corresponds well to the nonparametric fit from the locally-weighted polynomial regres-

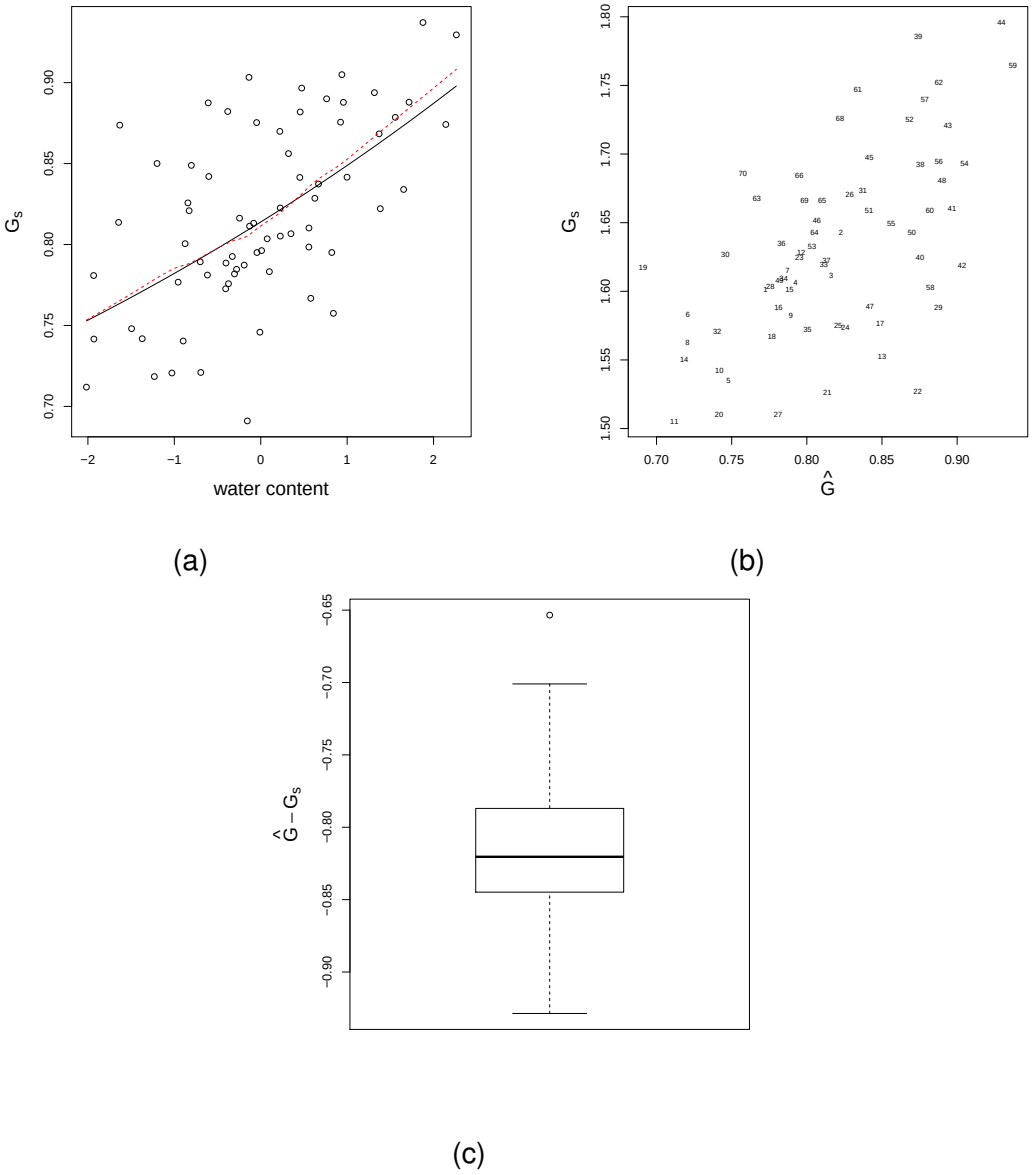


Figure 4.9: (a) the $\hat{G}$ as a function of water content, the dashed line corresponds to a nonparametric fit and the solid line represent the fit form Gini index model, (b) scatter plot of $\hat{G}$ versus G_s, the black solid line is the diagonal line, and (c) box plot of the difference between $\hat{G}$ and G_s.

sion (the dashed line). We check the model fit through Figure 4.9(b), in which the the nonparametric Gini index estimate is plotted against the predicted one resulting from model (4.31). Generally speaking, the proposed model has the capacity of detecting the underlying pattern in the observed data however, we see still large variability in the data. To the best of our knowledge, this is due to the complex data generating mechanism underlying and lack of other covariates as well. Besides, **?** suggested three

aspects of the orbited mites data set that may lead to the poor coverage of explained variation. Figure 4.9(c) shows the boxplot of the difference between $\hat{G}$ and G_s, it is skewed to the right and the median of the difference is close to zero.

4.8 Disscussions

Understanding how species biodiversity changes with respect to the environmental variables is crucial for classical biodiversity protection programs. In this chapter, we present a semiparametric model framework to model the Gini index as a function of a covariate. This model is only evaluated in the situation where species abundances follow a Poisson or a negative binomial distribution, but the developed method might also be applicable for data generated from other distributions. This, however, requires a careful evaluation of the appropriateness of the L2 model approximation.

The sandwich variance estimator for the L2 moment model can be extended to clustered data. The simulation studies indicate that the sandwich variance estimator overestimates for the i.i.d. setting and results in a conservative confidence interval, whereas, in the clustered setting, the sandwich variance estimator tends to underestimate, thus, lead to a liberal confidence interval. The sandwich covariance estimator and the estimator derived from the influence functions are worked out for estimating the covariance of the model parameters. The variance estimator derived from the influence functions is only valid for i.i.d. data. Since the data are clustered in the case study, we assessed the validation of the sandwich estimator by simulating clustered data. For a small within cluster variation, the sandwich estimator is approximately unbiased. For a larger within cluster variation, the bias increases. More researches need to be done for deriving the

influence function based variance estimator for the clustered data.

Although in this thesis the Gini index model is only well assessed for accommodating single continuous predictor, the proposed Gini index model can likely be extended to multivariate setting upon additional simulation studies. Besides, in this chapter, we proposed that the covariate should enter the model in the form of a summation ($\boldsymbol{X} + \boldsymbol{X}'$), this statement actually depends on the research context. For instance, one wants to learn how the genetic distance effect on the microbiodiversity, then the form of the covariate is more likely to be $\boldsymbol{X} - \boldsymbol{X}'$ where $\boldsymbol{X}$ denotes the DNA loci information. However, Model (4.23) is merely an approximation of the true L2 moment model. Its validity should be carefully evaluated via simulations that best mimic realistic data. In our simulation studies, species within a sample location are all generated with the same expected abundance. This may not be the best choice for reflexing the complex coexistence structure among the species since it assumes complete evenness on average. We also simplified Model (4.23) by replacing a double exponential by a single exponential and simulated responses from count models without a quadratic covariate, ignoring bell-shaped responses. More realistic settings would be valuable for assessing the empirical properties of the method, but this falls beyond the scope of this dissertation.

The proposed Gini index model is a new model, although we have studied the model intensively, there are still aspects that can be further studied. As we have addressed in Chapter 1, the biotic interaction has been ignored when developing the Gini index model. More complex model framework for better accounting for the correlation within organisms in the community can be developed. The inference of the influence function based variance estimator for the clustered design is absent, besides, a better option for the influence function of the L2 moment model should be further studied. The influence

function we introduced for the L2 moment model is just an arbitrary one, there likely exist more efficient choices. **?** has studied the efficient estimator for the probabilistic index model but the practical implementation remains a challenge.

In our treatment of the theory we further assume that abundances of the same species in different samples, say Y_{ik} and Y_{jk} ($i \neq j$), are independently distributed. However, when each species is modelled by a specific species response function, the pseudo-obervations $\mid Y_{ik} - Y_{il} \mid$ and $\mid Y_{jk} - Y_{jl} \mid$ are generally no longer independently distributed. This issue is ignored in the development of our theory. Future research could focus on solutions by e.g. integrating models for the species response functions into the Gini index model, or by accounting for the inter-sample dependence in the inferential procedures.

A case study is used to demonstrate model interpretation and diagnosis in Section 4.4. It is worth mentioning that when there are n abundances the computational burden goes up to n^2. Consequently, when our model is about to be applied to large data set, more efficient computing strategy, e.g., parallel computing may be desired.

Chapter 5

Human infant gut microbiome analysis

In this chapter, we analyse data from a study of the human infant gut microbiome [**?**]. The primary objective of the study is to examine the relationship between the human infant gut microbiome dynamics and type 1 diabetes. This chapter is organised as follows: in Section 5.1 an introduction to the data set is given and some data explorations are presented. In Section 5.2 we apply the constrained ordination analysis in the presence of zero inflation to study the microbiome composition with respect to food intake of infants. In Section 5.3 the Gini index model of Chapter 4 is used to examine the relationship between the microbiome diversity and age.

5.1 Data description

The infant gut microbiome data set comes from a longitudinal study of **?**. The original research objective was to identify the link between the development of the human infant gut microbiome and type 1 diabetes (T1D) in infants. Thirty-three infants predisposed

to T1D were included in the study and followed from birth until 3 years of age. Stool samples of each infant were taken on a regular basis to assess the composition of the gut microbiome. In total, 777 samples were collected and abundances of 2239 OTUs were obtained. Data on dietary intake were also recorded at each visit.

In this chapter, we analyse the microbiome data at family level: the 2239 OTUs are summarised into 48 microbial families. The OTUs with undefined family are left out of the analysis. The dimension of the abundance matrix $\boldsymbol{Y}$ is 777×48. Figure 5.1 shows the frequencies of zero abundances in the data set. The median of the number of zero abundance per family is approximately 400. Hence, about 50% of the abundance matrix consists of zeroes. When a child is diagnosed with T1D during the course of the study, it is considered as a T1D case.

Table 5.1 gives an overview of the 15 dietary intake variables, that are all binary, indicating whether a particular food product is part of the child's diet at the time of the stool sample collection. The age at the time of the data collection is also included in the data set (measured in days after birth). Table 5.2 is a frequency table of the dietary intake data, presented for five age categories. We can see that none of the infants older than 742 days receive breast feeding, whereas more than 78% of the infants younger than 251 days still receive breastfeeding. Confounding between age and breast feeding maybe a concern when interpreting our statistical data analysis. Figure 5.2 gives the box plots of the age at stool sample collection for each infant.

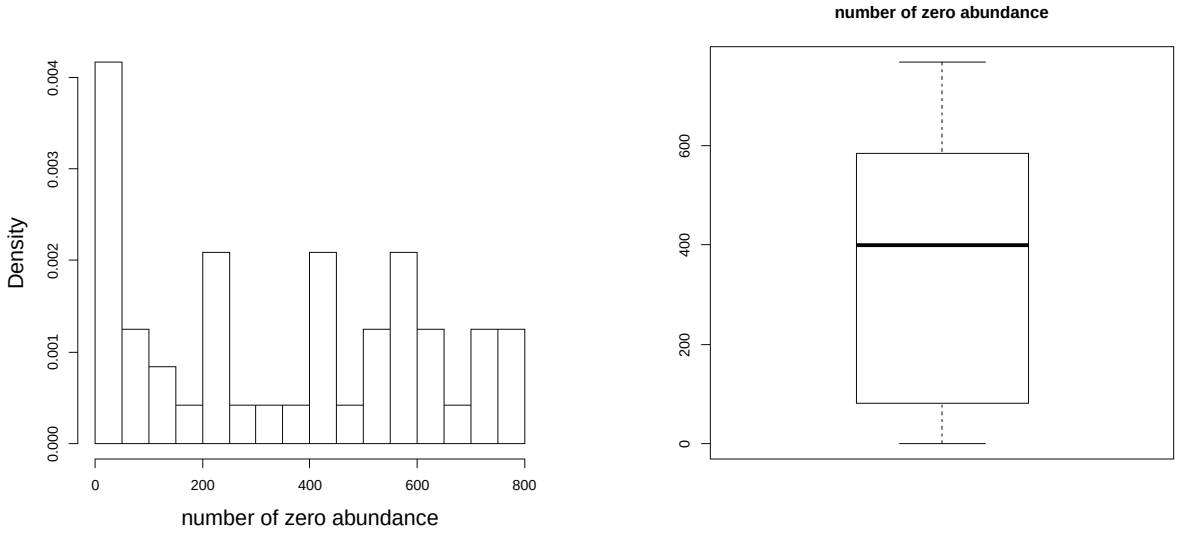


Figure 5.1: The histogram (left) and the boxplot (right) of the number of zero abundances per family.

Variable name	Description	Type and summary
BF	Breast feeding	binary: 248 yes and 529 no
Infant_Formula	Infant formula	binary: 663 yes and 114 no
Oat	Oat	binary: 638 yes and 139 no
Barley	Barley	binary: 531 yes and 246 no
Rye	Secale cereale	binary: 524 yes and 263 no
Root_Veg	Root vegetable	binary: 679 yes and 98 no
Buckwheat_Millet	Grain-like seeds: buckwheat and millet	binary: 198 yes and 579 no
Cereal	Cereal	binary: 664 yes and 113 no
Veg	Vegetable	binary: 649 yes and 128 no
Eggs	Eggs	binary: 477 yes and 300 no
Soy_Prod	Soy products	binary: 107 yes and 670 no
Milk_Prod	Milk products	binary: 560 yes and 217 no
Meat	Meat	binary: 637 yes and 140 no
Fish	Fish	binary: 581 yes and 196 no
Solid_Food	Solid infant food	binary: 681 yes and 96 no
Age_at_Collection	Age at collection (in days)	integer with average 482.8906 and standard deviation 294.7245

Table 5.1: Overview of the variables of infant dietary intake.

age (in days)	BF	Infant_Formula	Oat	Barley	Rye	Buckwheat_Millet	Cereal	Root_Veg	Veg	Eggs	Soy_Prod	Milk_Prod	Meat	Fish	Solid_Food
≤ 251	78.60	75.35	37.67	18.14	18.60	6.51	47.44	54.42	41.40	5.58	1.86	13.95	34.88	20.47	55.35
(251,497]	31.46	86.38	97.65	82.63	80.28	32.39	100.00	100.00	99.06	76.06	11.74	88.73	100.00	92.02	100.00
(497,742]	6.38	86.17	100.00	88.30	86.70	34.57	100.00	100.00	100.00	87.77	19.68	95.74	100.00	95.74	100.00
(742,988]	0.00	97.14	100.00	92.38	86.67	34.29	100.00	100.00	100.00	88.57	27.62	100.00	100.00	100.00	100.00
> 988	0.00	94.64	100.00	94.64	87.50	25.00	100.00	100.00	100.00	80.36	21.43	100.00	100.00	100.00	100.00

Table 5.2: For five age classes, the relative response for all binary dietary variables are listed.

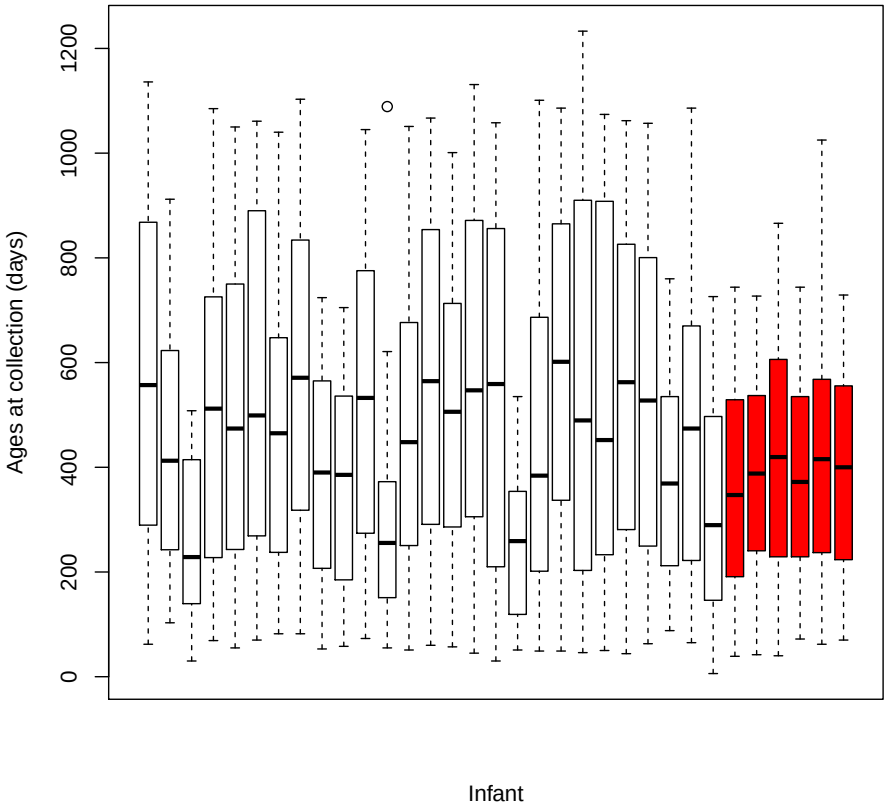


Figure 5.2: Box plots of the ages at stool sample collection for all 33 infants. The red boxes correspond to infants who developed T1D during the course of the study.

5.2 Constrained ordination analysis

Before proceeding, variable *Root_Veg* is removed from the analysis since it highly correlated with variable *Solid_Food* ($r = 0.99$).

A constrained ordination analysis may be helpful to gain insight into the variability of the abundance data and it may give a suggestion as to how the abundances vary with the dietary intake. This could give a first impression about the effect of the diet on the gut microbiome. Such an ordination analysis is often applied as a first data exploration

step because the two-dimensional ordination graphs are often easy to interpret.

Since the abundance data contain many zeroes the new constrained ordination method of Chapter 2 seems appropriate. Moreover, Figure 5.3 shows that the abundances are also overdispersed as compared to a Poisson distribution. In this section we apply two versions of our method: with the Poisson and with the ZINB distributions. For the purpose of benchmarking the analyses, we also analyse the data with the classical CCA method. Given the zero inflation and overdispersion of the data, our ZINB-based approach is expected to produce the most reliable results, because (1) the ZINB-based approach imposes no rigid equal tolerance assumption which is required by CCA; (2) the ZINB-based approach takes both zero inflation and overdispersion into account, which neither CCA nor Poisson-based methods are capable of. Another argument in favour of the model-based method is that the microbiome data are from 16S rRNA sequencing and the resulting abundances across samples are not comparable due to varying library sizes. In both the Poisson and ZINB-based approaches, we include the library size (i.e., the total number of counts in the sample) as an offset in the model to account for the library size variation. The classical CCA method, on the other hand, does not allow for an easy solution to this issue.

The abundance data is represented by the $n \times K$ matrix $\boldsymbol{Y} = \{y_{ik}\}$ which contains the abundances y_{ik} of micro-organism $k = 1, \ldots, K$ and with index $i = 1, \ldots, n$ referring to the $n = 777$ stool samples. It should be noted that the $n = 777$ stool samples are not independently sampled, because each of the 33 infants donated several stool samples. However, as an ordination method does not aim at performing statistical inference and only aims at exploring the data, the dependence in the data is no substantial issue in the application of the ordination methods. The index i thus refers to a unique combina-

tion of an infant and a stool sample collection, to which we refer as a *visit* or a *sample*. The $n \times p$ matrix $\boldsymbol{X} = \{X_{ij}\}$ contains the age at collection and the dietary observations x_{ij} of visit i on variable $j = 1, \ldots, p$ ($p = 15$). We use $\boldsymbol{x}_i$ to denote the i^{th} row of the dietary intake matrix $\boldsymbol{X}$. Our method results in estimating the vectors $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ by maximising the likelihood-ratio criterion (2.3). We refer to them as the first two *dietary gradients* (instead of *environmental gradients*). The linear combinations $z_{1i} = \boldsymbol{\alpha}_1^t \boldsymbol{x}_i$ and $z_{2i} = \boldsymbol{\alpha}_2^t \boldsymbol{x}_i$ $(i = 1, \ldots, n)$ are now referred to as the first two *dietary scores* of visit i. Finally note that we give names to $\boldsymbol{X}$ and $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ including the term "*dietary*", but also age at collection is included.

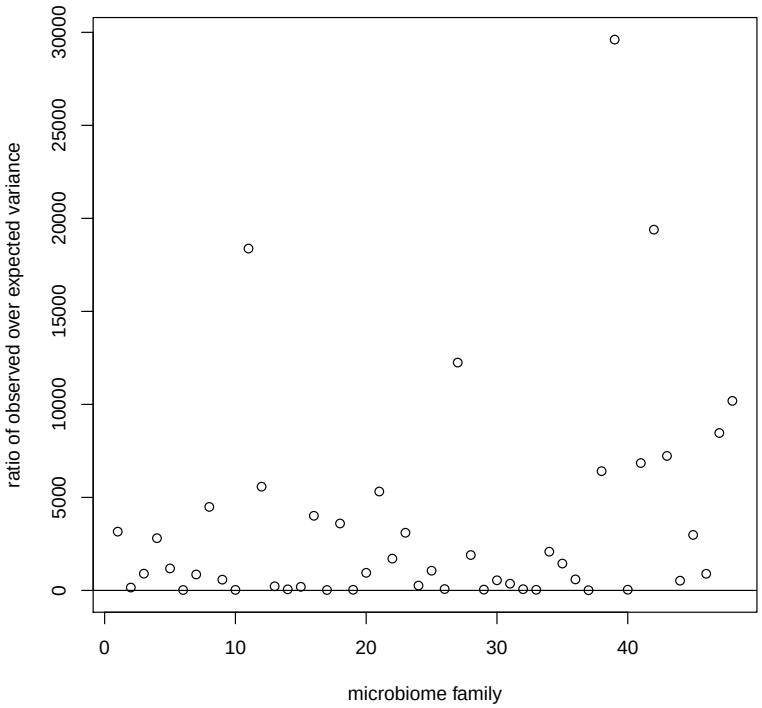


Figure 5.3: Overdispersion in the abundance data set. The ratio of the observed variance over the expected under the Poisson assumption is plotted for all microbiome families. The solid horizontal reference line represents a ratio of 1.

Table 5.3 shows the estimated first two dietary gradients from all three methods. From the results of the CCA we conclude that all the dietary variables are important in com-

posing the first ordination, but *infant formula*, *buckwheat millet* and *soya products* appear to be less important. The first CCA gradient shows a contrast between *breast feeding* and all the others. This can be interpreted as follows: the microbiome community at family level in the breastfed infants is different from those who are not breastfed. However, given the strong level of confounding between breast feeding and age, the first gradient may also interpreted in terms of the age of the child, i.e. non-dietary related changes in the child may also strongly affect the gut microbiome. In the second dimension the most important variables become *buckwheat millet* and *soy products*, which form a contrast.

The Poisson-based approach produces quite different results compared to the CCA. This can be explained by the fact that the response functions are no longer restricted to have equal tolerances. From the estimates of the dietary gradients, we conclude that the most important variable in the first dimension is *Soy products* while *Rye* and *age* become the most influential variables in the second dimension.

For the ZINB-based approach, the key variables in composing the first ordination are *eggs*, *soy products*, *milk products* and *breast feeding*. Moreover, *breast feeding* contrasts with the other three. The interpretation is that the microbiome in the gut of breastfed infants is different from the microbiome of infants who are not breastfed but receiving eggs, soya products and milk products. In the second dimension *buckwheat millet*, *eggs* and *age* are of most importance. As before, care must be taken when interpreting the effects of age and breast feeding because of their very strong association. The different result obtained based on ZINB compared to the one based on Poisson is caused by the excess zero abundances and overdispersion in the data.

The three methods give different interpretations to their first two gradients. However, for each method the two gradients only span a 2-dimensional subspace, and despite differences in the two gradients between methods, the 2-dimensional subspaces may still be similar between methods. In other words, the three sets of two gradients may be similar up to a orthogonal transformation.

	CCA		Poisson		ZINB	
	α_1	α_2	α_1	α_2	α_1	α_2
BF	-0.7059	0.2112	-0.1623	-0.0031	-0.3515	0.1486
Infant_Formula	0.2705	-0.1375	0.1985	-0.0772	0.2151	0.0157
Oat	0.7934	-0.1427	0.2234	-0.0123	0.2547	0.2124
Barley	0.6050	0.2032	0.0812	-0.1962	-0.0397	0.2884
Rye	0.5192	0.1210	-0.1885	0.6367	-0.1689	0.2281
Buckwheat_Millet	0.1916	0.4024	-0.0543	0.0939	-0.0380	0.4235
Cereal	0.7111	-0.0434	0.2039	-0.0188	0.1190	0.0789
Veg	0.7374	-0.0464	-0.0093	0.2052	-0.0642	0.0515
Eggs	0.7760	0.1659	0.3614	0.2631	0.4107	0.4334
Soy_Prod	0.2636	-0.4942	0.6312	-0.0391	0.4545	-0.1416
Milk_Prod	0.8087	0.0150	0.3701	-0.0616	0.5532	0.3486
Meat	0.7672	-0.0626	-0.2174	-0.2571	-0.0116	0.2525
Fish	0.8242	0.0569	0.1034	0.0174	-0.0158	0.1526
Solid_Food	0.6545	-0.0200	-0.1385	-0.0206	0.1007	-0.0037
Age_at_Collection	0.8097	0.0381	0.2334	0.5978	0.1541	0.4388

Table 5.3: The estimated first two dietary gradients from CCA, and the Poisson and ZINB-based approaches.

To have an idea of how dissimilar the estimated dietary gradients are, the Grassmannian distances [**?**] between the three 2-dimensional spaces spanned by the first two dietary gradients from three different approaches are calculated. The distance matrix is given below, and we use α_c, α_p and α_z to denote the two-dimensional spaces from CCA, the Poisson and ZINB-based approaches, respectively:

$$\begin{array}{c} \\ \alpha_c \\ \alpha_p \\ \alpha_z \end{array} \begin{array}{c} \begin{array}{ccc} \alpha_c & \alpha_p & \alpha_z \end{array} \\ \begin{pmatrix} 0.0000 & & \\ 1.496 & 0.0000 & \\ 0.9008 & 1.2105 & 0.0000 \end{pmatrix} \end{array}$$

We conclude that the three subspaces spanned by the environmental gradients are all quite dissimilar. We therefore expect to find different conclusions from the three ordination plots.

Ordination diagrams for the three approaches are presented in Figures 5.4. The samples are represented by circles, the microbiome families by plus signs, and arrows represent the dietary variables. From the left panel of Figure 5.4 we see a clear arch effect [**?**], which is a well known phenomenon for CCA. The arch effect is not present in the other graphs. Interestingly, the middle and right panel of Figures 5.4 reveal another structure: clusters of samples lie along smooth lines. The explanation is that the data from the same infant do not vary much over time. In particular, dietary intake changes only slowly with age (size of the circle). For example, most of the observations from subject E003872 lie along a straight line, and the same holds for subjects E001463 and E003989. Figure 5.5 illustrates that in the early days of an infant's life, the scores

of subject E001463 vary strongly with age, but from a certain age onwards, the scores on the two dimensions seem to correlate with each other in almost a linear way. By checking Table B.7 in Appendix B.4 we arrive at the explanation: the dietary intake of this infant varies strongly in the first year, but from 366 days onwards, the dietary intake stays stable. From the right panel of Figure 5.4, We observe that most of the circles which lie in a straight line are from the same subject, and we come to the conclusion that for many infants, the composition of the microbiome shifts slowly and steadily over time. We categorise the age into five levels and represent it by the size of the circles. It is interesting to see that the microbiome composition is rather consistent for the older infants (older than 742 days). The microbiome composition varies substantially for the younger infants.

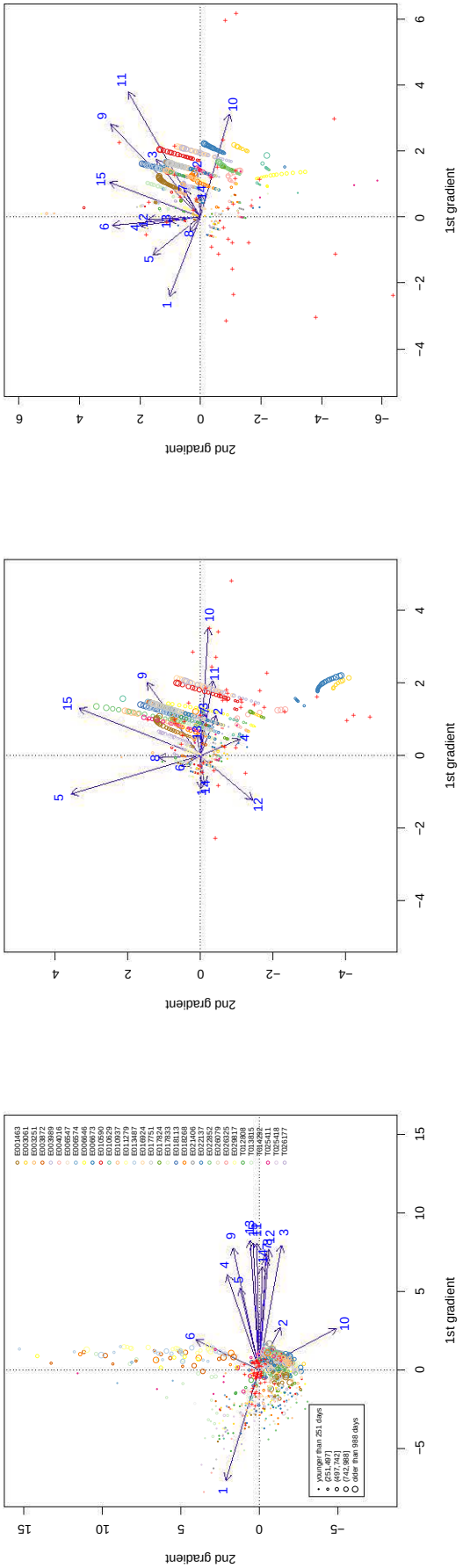


Figure 5.4: Ordination diagram for the human infant gut microbiome data analysed with CCA (left), the Poisson-based approach (middle) and the ZINB-based approach (right). Circles represent samples, plus signs represent families, arrows represent the dietary intake variables, the size of the circles is proportional to the age, and the subjects are annotated by colour. The dietary variables: 1, *BF*; 2, *Infant_Formula*; 3, *Oat*; 4, *Barley*; 5, *Rye*; 6, *Buckwheat_Millet*; 7, *Cereal*; 8, *Veg*; 9, *Eggs*; 10, *Soy_Prod*; 11, *Milk_Prod*; 12, *Meat*; 13, *Fish*; 14, *Solid_Food*; 15, *Age_at_Collection*.

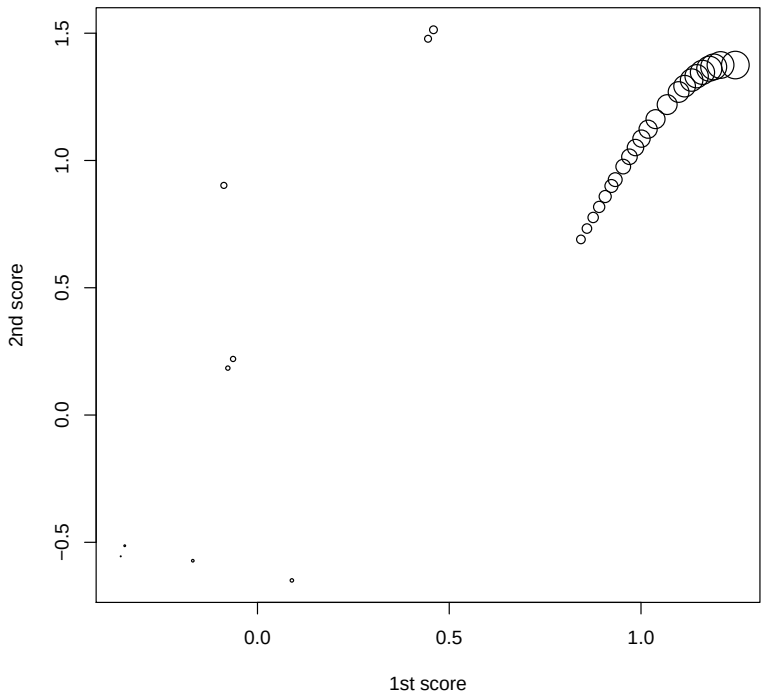


Figure 5.5: The sample scores of subject 'E001463'; the size of the circles is proportional to the age of the infant.

The three plots in Figure 5.4 more or less all imply that the micriobiome composition in the young infants (younger than 251 days) is quite different compared to the older infants (older than 742 days).

To demostrate whether the ZINB-based approach is indeed more informative for the human infant gut microbiome data, we focus on the microbiome family *Veillonellaceae*. The reason we choose *Veillonellaceae* is that it is the most abundant microbiome family in the infants younger than 251 days. Five samples are also highlighted to assist the comparison of the results. From Figure 5.6, we conclude that *Veillonellaceae* appears to be closer to the samples from the younger infants from the ordination diagram based on Poisson and ZINB, whereas, from the CCA ordination digram this can not be concluded. For the left panel of Figure 5.6 we observe that the *Veillonellaceae*

should be most abundant in samples G36048 or G36534, whereas for the Poisson-based approach, G37018 should be the optimum and for the ZINB-based approach, the optimum is rather G37049. By checking the observed abundances of the *Veillonellaceae* in those aforementioned samples (Table 5.4), we conclude that in G37049 the *Veillonellaceae* are most abundant. Thus, the result from the ZINB-based approach corresponds to the reality in this example. Next we try to draw some insightful conclusions about the relationship between the microbiome composition and the dietary intake, based on the complete data set.

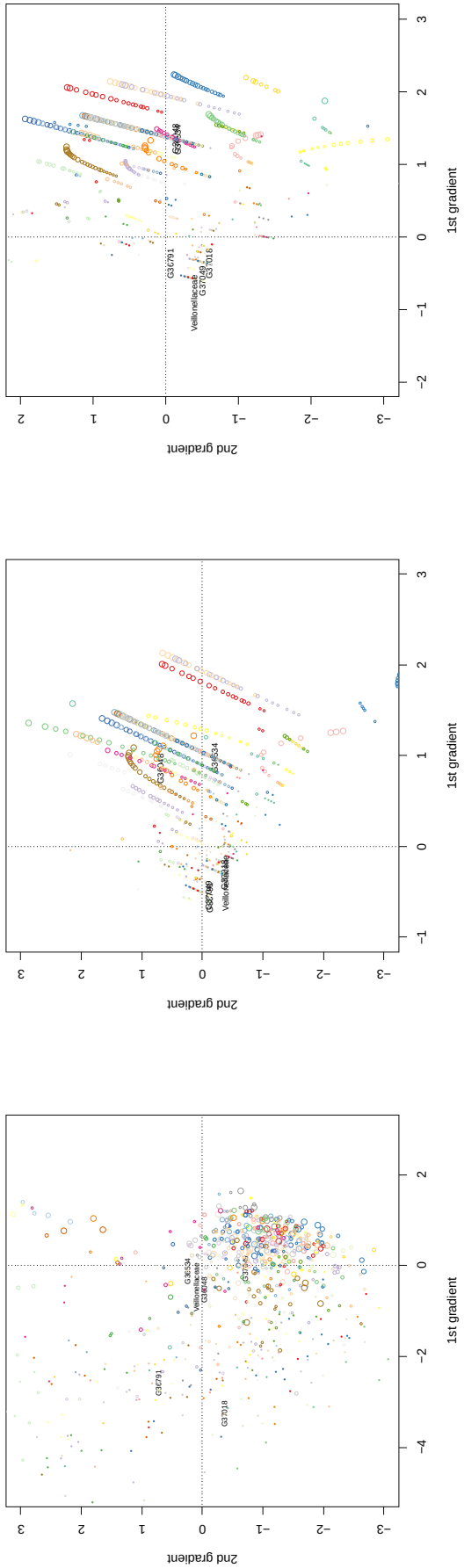


Figure 5.6: Ordination diagram for the human infant gut microbiome data analysed with CCA (left), the Poisson-based approach (middle) and the ZINB-based approach (right). Circles represent samples, plus signs represent families, the size of the circles is proportional to the age, and the subjects are annotated by colour (see the legend in the left panel of Figure 5.4). To increase the readability we only indicated the *Veillonellaceae* family, and only 5 samples are indicated with their ID code.

sample	G36048	G36534	G37049	G36791	G37018
abundance	12	456	1378	63	539

Table 5.4: The observed abundance of the *Veillonellaceae* in samples G36048, G36534, G37049, G36791 and G37018.

In the ordination diagram of Figure 5.7 the T1D cases (green circles) are distinguished from the infants who did not develop T1D by the end of infancy (black circles). It can be seen that the samples from the T1D cases are more shifted to the left on the first ordination axis, and the non-T1D samples are found more to the right, but there is some overlap between the two groups. The ordination diagram suggests that T1D cases are more breastfed, eat more rye and less soy products than the other infants. The non-T1D cases seem to eat more soy and milk products, eggs and infant formulas than the T1D cases. Despite the strong correlation between age and breast feeding in the data, which warrants care in interpreting separate effects of breast feeding and age, the second ordination dimension correctly suggests that more diabetes is observed at the later ages of the infants included in the study. Also note that the age and breast feeding directions are almost orthogonal in the graph. This feature, however, is not implied by the method.

In Appendix C we provide the ordination diagrams with samples distinguished according to (1) each diet variable and (2) whether or not the infant developed T1D.

5.3 The Gini index model

In this section we will model the Gini index of the human infant gut microbiome data as a function of age using the Gini index model (4.31) developed in Chapter 4. We

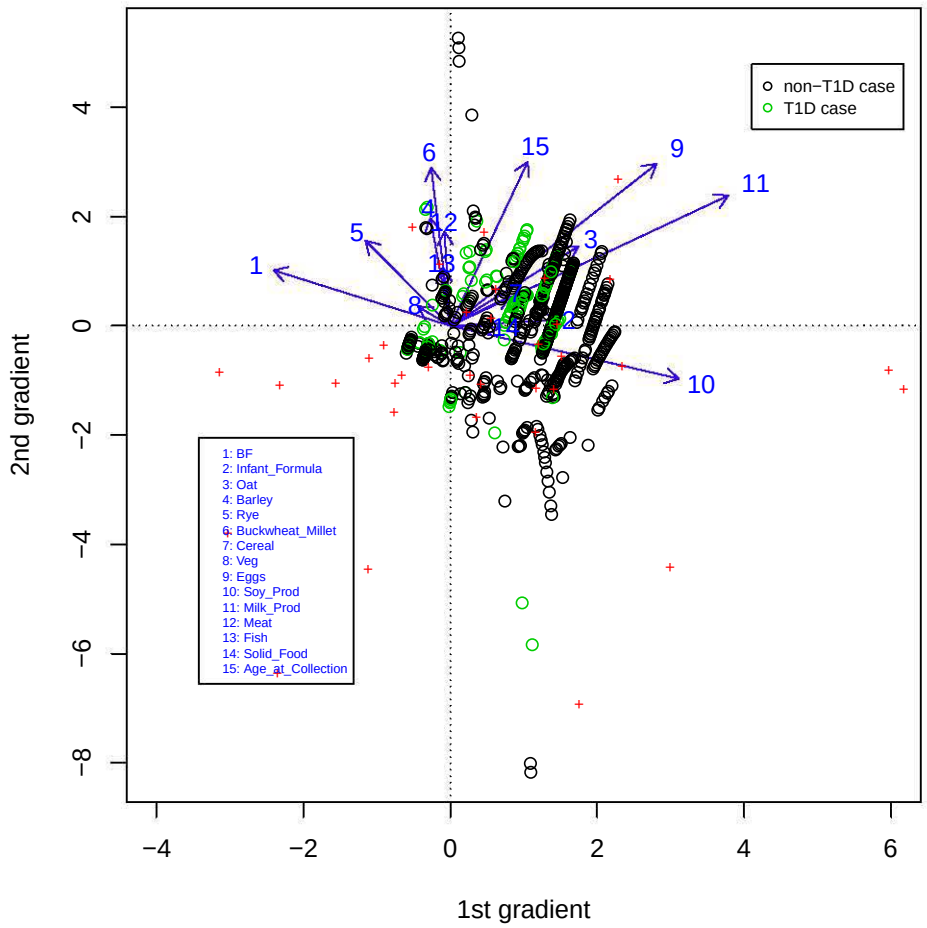


Figure 5.7: Ordination diagram for the human infant gut microbiome data analysed with the ZINB-based approach. Arrows represent the dietary intake of the infants, plus signs represent families, and circles represent the samples. The samples are coloured according to status of T1D.

restrict the data analysis to a univariate setting. The multivariate setting is beyond the scope of this thesis. The objective is to estimate how the microbiome diversity varies with age, and to study the size and significance of the association. Let X_i denote the age of the child at collection of sample i. Let $G(X_i)$ denote the expected Gini index of the microbiome of a child of age X_i. Let Y_{ik} denote the abundance of microbial family k in sample i.

Recall that the new Gini index model involves an L1 moment model and an L2 moment model. In particular,

$$\begin{aligned} 2G_s(X_i) &= \frac{\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i\right)}{\mathrm{E}\left(Y_{ik} \mid X_i\right)} = \frac{\exp\left(\beta_0 + \beta_1(X_i + X_i) + \beta_2(X_i^2 + X_i^2)\right)}{\exp\left(\alpha_0 + \alpha_1 X_i + \alpha_2 X_i^2\right)} \\ &= \exp\left(\beta_0 - \alpha_0 + (2\beta_1 - \alpha_1)X_i + (2\beta_2 - \alpha_2)X_i^2\right) \\ &= \exp\left(\gamma_0 + \gamma_1 X_i + \gamma_2 X_i^2\right), \end{aligned} \tag{5.1}$$

where $\gamma_0 = \beta_0 - \alpha_0$, $\gamma_1 = 2\beta_1 - \alpha_1$ and $\gamma_2 = 2\beta_2 - \alpha_2$ and where $k < l$. The original micriobiome abundance data set contains abundances of 48 microbial families in 777 samples. These samples are taken from 33 distinct infants at different time points. We will, however, fit the model to only a subset of the data. This is because of the following two reasons:

- the theory in Chapter 4 is established for the i.i.d setting or clustering per infant for a fixed time point. The human infant gut microbiome data, however, are from a longitudinal study: repeated measurements on each infant are taken at different time points. This makes the estimation procedure of the L2 moment model complicated. Therefore we will only consider a cross-sectional part of the study.

- for n abundances, the L2 moment model requires $\frac{n(n-1)}{2}$ pseudo-outcomes. For the entire microbiome data set, this means using more than 6.8283×10^6 pseudo-outcomes. Currently the computational cost is too high for our algorithms.

To cope with these two aspects, we continue with a cross-sectional subset of the data. This subset is constructed such that only one sample from each infant is randomly selected. The number of samples in the subset is thus 33. Figure 5.8 shows the box plots of age in the original data set and in the subset. By selecting a subset of the data as described, the study becomes cross-sectional and the computational complexity is decreased to a great extent.

Note that even the cross-sectional subset does not result in independent data since for each child we have abundances for multiple microbial families. To account for this clustering, we estimate the variances of the regression parameters via bootstrapping (1000 runs). In Chapter 4 we also argued that even dependence between independent biological samples may arise because of expected abundances shared by identical species. The latter type of dependence should be studied in further research and is ignored in our data analysis.

5.3.1 The L1 moment model

Consider the following random intercept L1 moment model ($i = 1, \ldots, 33$, $k = 1, \ldots, 48$)

$$\mathrm{E}\left(Y_{ik} \mid b_i, X_i\right) = \exp\left(b_i + \alpha_0 + \alpha_1 X_i + \alpha_2 X_i^2\right), \tag{5.2}$$

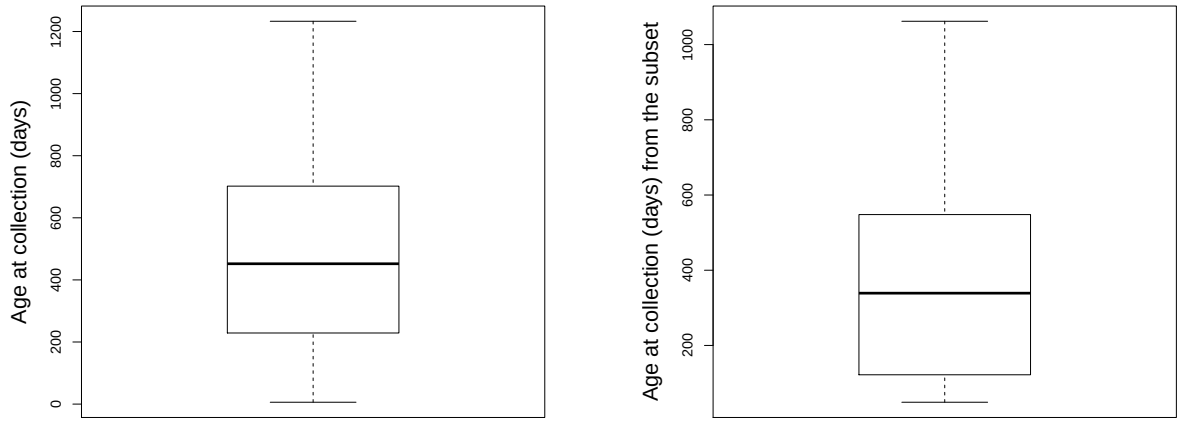


Figure 5.8: Box plots of the age from the entire data set (left) and from the selected subset of the data (right).

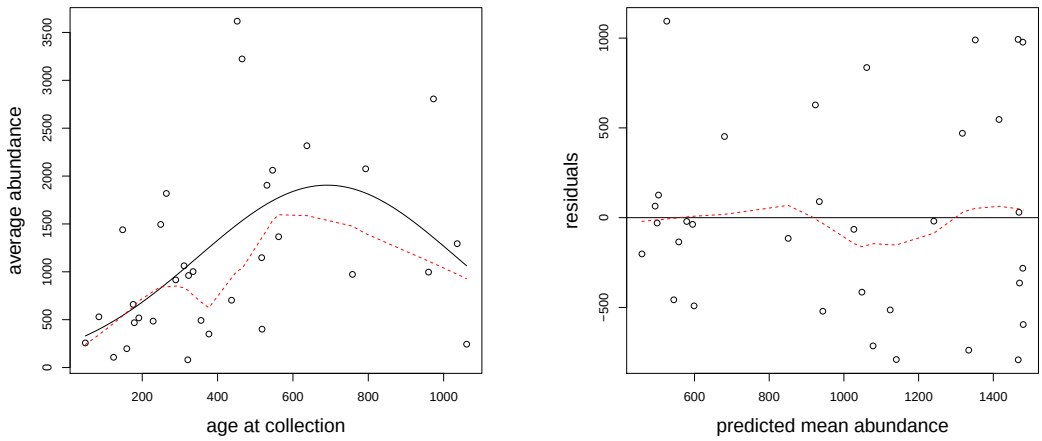
where b_i a sample-specific random effect that follows a normal distribution and is independent from X_i. In Table 5.5 the estimates of the regression coefficients are given together with their standard errors. The estimates are obtained by solving the estimating equation (4.3) (on page 112). The standard errors result from the bootstrap procedure described earlier. From the p-values we see that age has a significant effect on the mean abundance at the 5% significance level.

Figure 5.9(a) shows the per-sample average abundance as a function of age. From this figure and from Table 5.5 we conclude that the quadratic term of age is not significant at the 5% level of significance; the quadratic effect seems to be caused by the observations of the oldest infants in the subset. A wald type test is applied to compare model (5.2) with an intercept-only L1 moment model, resulting is p-value of 0.00003. Thus, we conclude that although the quadratic term is not statistically significant, the age still shows significant effect on the mean abundance. From the method development of Chapter 4 we know that the quadratic effect needs to be included for the L2

moment model, and thus in the final Gini index model. The results indicate that a large fraction of the between-sample variability remains unexplained. The residuals plot in Figure 5.9(b) shows no systematic pattern in the residuals, indicating the residuals are independent of the predictions resulting from model (5.2).

Table 5.5: Parameter estimates of the L1 moment model (5.2). The standard error is the bootstrap standard error.

	Estimate	Standard error	p-value
α_0	7.2070	0.1809	< 0.0001
α_1	0.5830	0.1681	0.0005
α_2	-0.2030	0.1593	0.2021



(a) (b)

Figure 5.9: (a) The average observed abundance as a function of age. The solid curve is the fitted curve from model (5.2) and the dashed line is from a nonparametric fit. (b) The residuals as a function of the predicted mean abundance. The dashed line is from a nonparametric fit.

5.3.2 The L2 moment model

Consider fitting the following L2 moment model to the data

$$\mathrm{E}\left(|Y_{ik} - Y_{il}| \mid X_i\right) = \exp\left[\beta_0 + \beta_1(X_i + X_i) + \beta_2(X_i^2 + X_i^2)\right], \tag{5.3}$$

where $i = 1, \ldots, 33$, $k, l = 2, \ldots, 48$ and $k < l$. The parameter estimates are given in Table 5.6, together with the bootstrap standard errors. The quadratic term is not significant at the 5% level of significance. We again apply a wald type test to compare model (5.3) with a L2 moment model with only an intercept term, resulting in a p-value of 0.0705. Hence, we conclude that age does not have significant effect on the L2-moment at the 5% level of significance. The p-value is however close to the level of significance. In Figure 5.10(a) we show the per-sample average pseudo-outcomes as a function of age. The i^{th} mean observed pseudo-outcome is defined as $\sum_{k<l} \sum_{l=2}^{48} |Y_{ik} - Y_{il}|/1128$. Figure 5.10(b) shows the residual plot. From both plots we see that age does not explain all the variation in the pseudo-outcomes.

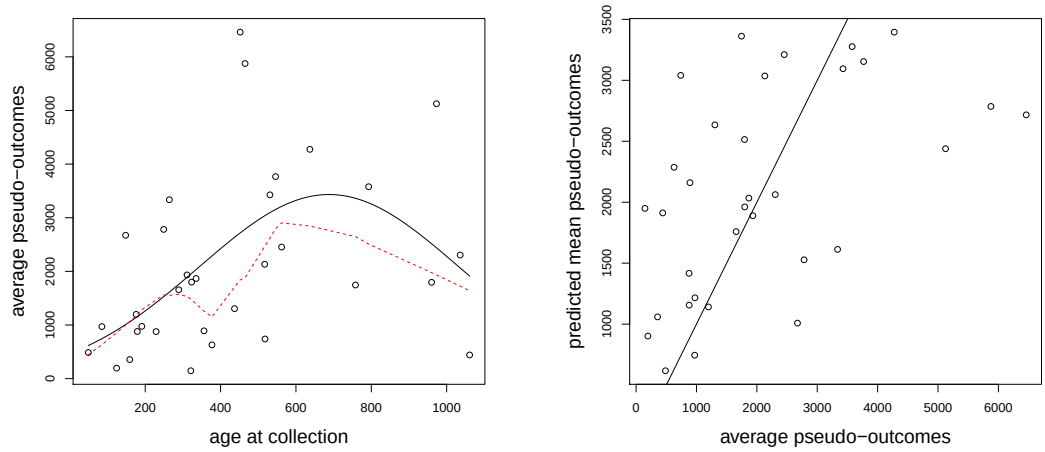
Table 5.6: Parameter estimates of the L2 moment model (5.3). The standard error is the bootstrapped standard error.

	Estimate	Standard error	p-value
β_0	7.7985	0.1791	< 0.0001
β_1	0.2805	0.0839	0.0008
β_2	-0.0986	0.0794	0.2144

Note that Figure 5.10(a) is remarkably similar to Figure 5.9(a). Figure 5.11 shows the relationship between the per-subject average abundances and the per- subject average pseudo-outcomes. A very strong positive correlation is observed.

5.3.3 The Gini index model

In Section 4.5 we have proposed an estimator of the variance of $\hat{\gamma}$ in model (5.2). This estimator is approximately unbiased for moderately clustered data. To see if this estimator is applicable here, we first check the degree of clustering of the data by fitting



(a) (b)

Figure 5.10: (a) The per-subject average pseudo-outcome as a function of age. The solid curve is the fitted curve from model (5.3) and the dashed line is from a nonparametric fit. (b) The residuals of the fit of model (5.3) as a function of the predicted mean pseudo-outcomes.

a random intercept model:

$$\mathrm{E}\left(Y_{ik} \mid b_i, X_i\right) = \exp(\alpha_0 + b_i + \alpha_1 X_i),$$

where b_i is a sample-specific random effect and X_i is the age. The estimated standard deviation of the distribution of b_i is 0.763. From Section 4.6.2, we know that the influence function based variance estimator exhibits finite sample bias when the standard deviation of the random intercept exceeds 0.5. To circumvent this we therefore propose the following bootstrapping procedure: for each bootstrap iteration, randomly sample the locations with replacement and fit model (5.2) to obtain $\hat{\gamma}$. We use the empirical variance of the $\hat{\gamma}$'s for inference.

Table 5.7 shows the fit of the Gini index model. The p-values show that age is not significant at 5% significance level. Figure 5.12(a) shows the nonparametric Gini index estimates (equation (4.1)) plotted as a function of age. In Figure 5.12(b) the model-

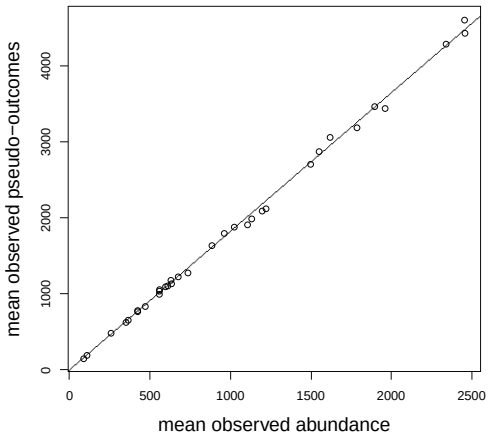


Figure 5.11: The relationship between the per-subject average abundances and the per-subject average pseudo-outcomes. The solid line is of the fit of the two response variables from a linear model.

.

based Gini index results from model (5.2). We see that the model-based Gini index does not correspond to the nonparametric estimates of the Gini index. The poor fit may result from the fact that the proposed Gini index model (4.31) is merely an approximation. Recall that the expression from **?** (equation (4.9)) involves a double exponential function, whereas in our approximation we only use a single exponential. Moreover, the terms A and B from equation (4.9) are not taken into account.

Figures 5.12(a) and (b) suggest that the microbiome diversity decreases during infancy, but a remarkable amount of between-subject variability remains. Figure 5.12(c) shows a side-by-side comparison of the nonparametric estimates of the Gini index and the model-based Gini index from model (5.2) and Figure 5.12(d) is the box plot of the difference between these two estimates. The range of the nonparametric estimates is wider than the one resulting from model (5.2) and the median of the difference is different from zero.

Table 5.7: Parameter estimates for the Gini index model (5.2). The standard errors are the bootstrapped standard errors.

	Estimate	Standard error	p-value
γ_0	0.5915	1.9626	0.7631
γ_1	-0.0220	0.0709	0.7563
γ_2	0.0058	0.7109	0.9935

Analysis from the conventional approach

We mentioned in Chapter 4 that the conventional data analysis method is a two-step approach: first, the Gini index is estimated nonparametrically as in equation (4.1). This is denoted by $\hat{G}$. Then a generalised linear model with a log link function is used to link $\hat{G}$ to the covariate, i.e.,

$$\mathrm{E}\left(\hat{G}_i \mid X_i\right) = \exp(c_0 + c_1 X_i + c_2 X_i^2), \tag{5.4}$$

where c_0, c_1 and c_2 denote the model parameters. A log link is applied to ensure positive fitted Gini index. Here we use generalised least squares with the estimated variance of $\hat{G}_i$ as the weight for the parameter estimation. The parameter estimates are given in Table 5.8. The p-values suggest that age does not have significant effect on the $\hat{G}$, this corresponds to the conclusion made based on the model semiparametric Gini index model. However, Figure 5.13 shows that the estimated Gini index from model (5.4) does not agree well with the nonparametric Gini index estimate.

	Estimate	Standard error	p-value
c_0	-0.0897	0.0738	0.2243
c_1	-0.0144	0.0564	0.7986
c_2	0.0044	0.0545	0.9353

Table 5.8: The parameter estimates of model (5.4)

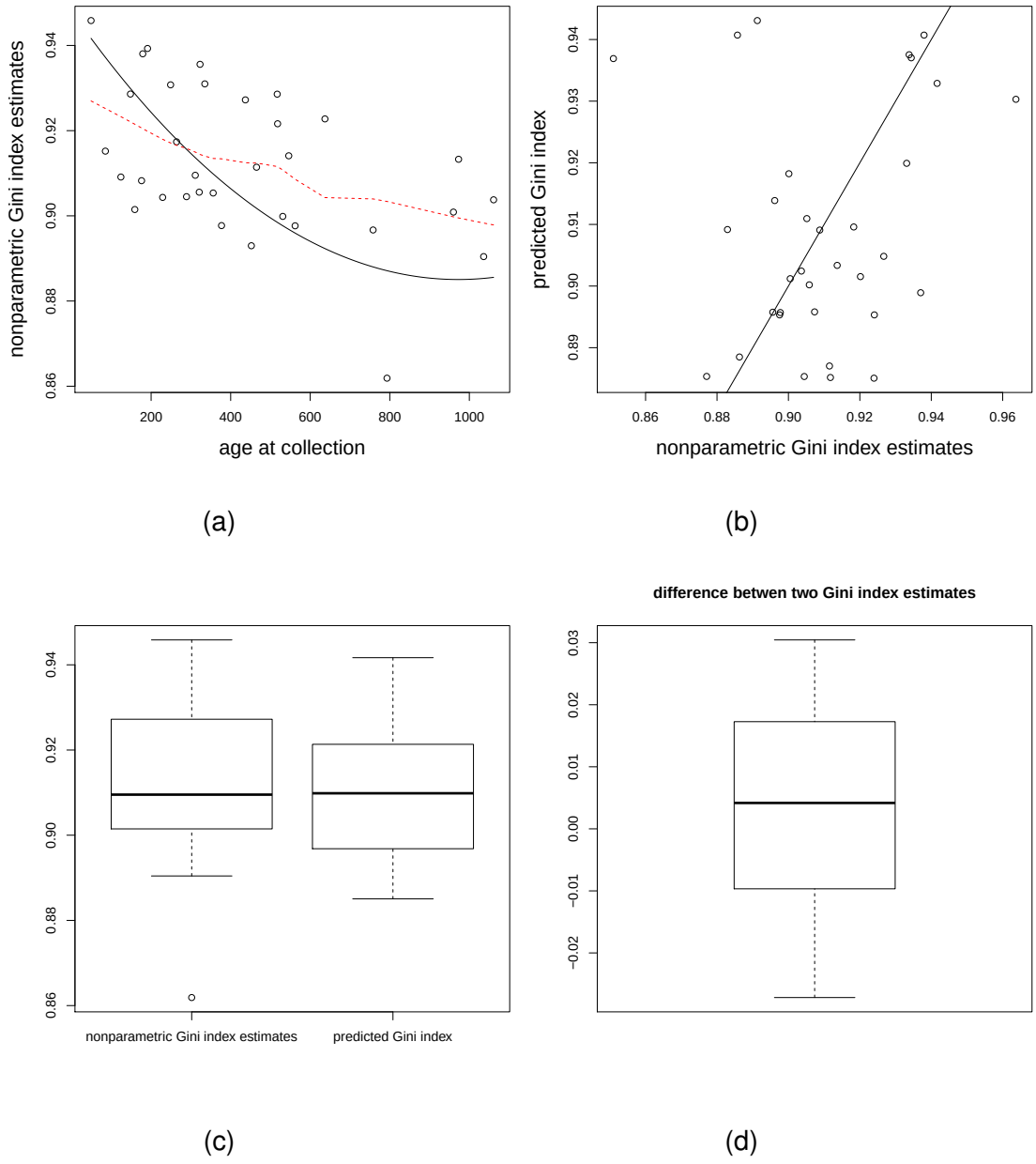


Figure 5.12: (a) The nonparametric estimates of the Gini index versus age. The solid line is the fit from model (5.2) and the dashed line represents the nonparametric fit using locally-weighted polynomial regression with 0.5 smoother span. (b) The model-based Gini index from model (5.2) plotted against the nonparametric estimate. (c) Box plots of the nonparametric estimates of the Gini index and the model-based Gini index. (d) Box plot of the difference between the nonparametric estimates of the Gini index and the model-based Gini index.

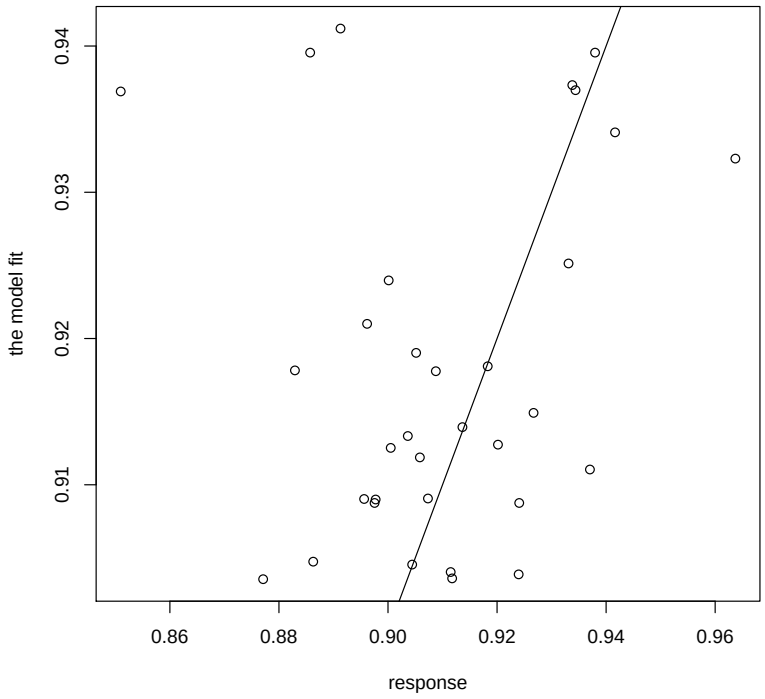


Figure 5.13: The response ($\hat{G}$) as a function of the model fit from (5.4). The solid line is the diagonal line.

Chapter 6

Conclusion and discussion

The body of this thesis consists of four individual research chapters: Chapters 2, 3, 4 and 5. Chapters 2 and 3 focus on modifications of existing ordination methods used in ecology. In Chapter 2, the constrained ordination analysis with flexible response function (FCOA) is modified to cope with zero-inflation and over(under)dispersion in abundance data. Chapter 3 is essentially a sequel to Chapter 2. In Chapter 3 we pointed out the fundamental ecological assumption is not guaranteed in FCOA and we therefore proposed a way to address this issue. In Chapter 4 we focused on modelling biodiversity as a function of a covariate using a semiparametric model. Chapter 5 presents an application of the methods developed in Chapters 2 and 4 to a human infant gut microbiome study. In the remainder of this concluding chapter, we present for each individual chapter a brief summary along with some future research perspectives.

In Chapter 2 we have first shown how ignoring the presence of excess zero abundance can lead to misleading results and error-prone conclusions. Zero inflation can have different causes in species-abundance studies. We improved the model-based

constrained ordination methods by replacing the Poisson distribution with hurdle zero-truncated Poisson (HZTP), hurdle zero-truncated negative binomial (HZNB), zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB). Both the zero-inflated models and the hurdle models have a parameter that refers to the presence of excess zeroes. This parameter may also be modelled through a logistic regression model with the environmental scores as predictors. For the hurdle models this parameter can be estimated separately and thus they have a computational advantage. Our methods were evaluated through simulation studies and were applied to real case studies. In addition, we also demonstrated how to obtain environmental gradients in higher dimensions so as to enable a two-dimensional graphical display of the results. However, two future research aspects of our proposed approach can be directly seen:

- Due to the zero-inflation structure, we often consider the following distributional assumption. For $i = 1, \ldots, n; k = 1, \ldots, K$

$$Y_{ik} \mid \boldsymbol{x}_i \sim Y_{ik} \mid z_i \sim \mathcal{M}_p,$$

  where $\mathcal{M}_p$ is a class of parametric models indexed by a (conditional) mean, say μ_{ik}, and a parameter that addresses the zeroes, say π_{ik}. We have only considered parametric models for μ_{ik} and π_{ik}, but they can be extended to generalised additive models (GAMs) [**??**]. **?** has generalised GAMs to include the zero-inflated exponential family. It requires linking a smooth function to π_{ik} and linking a smooth function to the μ_{ik} from the non-zero-inflated part. The complexity of the zero-inflated GAM lies in the fact that sometimes the data generating process of nonzero responses and that of the zero responses are not mutually independent.

? proposed a method called *constrained zero-inflated GAM* to solve the problem. Further efforts could include incorporating the zero-inflated GAM into Zhu's log-likelihood ratio criterion [**?**] and producing informative graphs. Computational issues are to be expected, since the procedure involves iterative maximisation for fitting the smooth regression function and for estimating the environmental gradient by maximising the log-likelihood ratio criterion.

- In order to search for environmental gradients in higher dimension, we simply forced the environmental scores to be orthogonal. This algorithm is practical but greedy. Another strategy that does not require orthogonality could be developed along the lines of the exploratory projection pursuit procedure proposed by **?**.

In Chapter 3 we have modified FCOA to increase the number of meaningful bell-shaped species response curves by introducing a penalty term in the estimation procedure. The method is named *bell-shape enriched COA* (BECOA). We illustrated our approach from a Bayesian point of view and showed that the penalisation is equivalent to introducing a normally distributed prior for which the mean must be negative so as to favour bell-shaped species response curves. We proposed two alternative algorithms for parameter estimation. We suggested determining the tuning parameter involved in the penalty term by controlling a trade-off between the goodness-of-fit of the model and the separation between the species response curves. The method involves 10-fold cross-validation. The approach was evaluated in a simulation study and applied to a real case study. At the end of the chapter we extended the method to bell-shaped enriched constrained ordination analysis for absence/presence data. For this purpose, we employed logistic regression for which we encountered the separation problem. We suggested to use Jeffreys invariant prior to avoid the difficulty of separation.

We briefly mentioned at the end of Chapter 3 that the zero-altered models from Chapter 2 can also be integrated into the bell shape enriched constrained ordination analysis (BECOA). For example, the Poisson density function may be replaced by HZTP density function. The penalised score vector (3.4) then becomes

$$\sum_{i=1}^{n} \frac{\partial \log p_k(\boldsymbol{\beta}_k \mid y_{ik}, z_i)}{\partial \beta_{jk}} = \sum_{i=1}^{n} \left(y_{ik} - \frac{e^{e^{(\boldsymbol{\beta}_k^t \boldsymbol{w}_i)}} e^{(\boldsymbol{\beta}_k^t \boldsymbol{w}_i)}}{e^{e^{(\boldsymbol{\beta}_k^t \boldsymbol{w}_i)}} - 1} \right) w_{ij} + n \frac{\partial \log g(\boldsymbol{\beta}_k)}{\partial \beta_{jk}} = 0.$$

Finding the solution of the penalised score equation above is not as simple as solving (3.4) since it is not linear in $\boldsymbol{\beta}_k$. However, one can use Newton-Raphson to find the solution. The score vector for maximising the LLR should be modified accordingly to accommodate the HZTP density function, and again Newton-Raphson can be employed for the maximisation procedure. Analogously, adopting BECOA to HZNB, ZIP and ZINB is only a matter of rewriting the penalised score vector and the expression of the LLR.

One obstacle we can foresee of the BECOA is again the computational burden, and this for several reasons. First, for data from metagenomics studies, which could involve hundreds to hundreds of thousands of taxa. Second, the approach itself involves iterative steps for maximisation. Third, the cross-validation for determining the tuning parameter can slow down the procedure significantly. Parallel computing could be part of the solution. There exist several packages in R that allow for parallel computing. We used *sfCluster*; see **?** for more details.

In Chapter 4 we illustrated that a semiparametric regression model can be used to describe the Gini index by relating the Gini index to the first two L-moments. It is conventional to use GLM for modelling the L1-moment, thus we focused on developing the

model framework for L2-moment. Next, we developed the semiparametric theory for L2 moment model relying on the sparse correlation theory [**?**]. We have shown that the resulting parameter estimator is asymptotically normally distributed and worked out a sandwich estimator for the covariance matrix. The proposed model is later extended to work for clustered designs. Next the need for deriving an estimator of the covariance between parameter estimators of both L-moment models has motivated us to pursue the L2 moment model in a semiparametric setting. We showed how to obtain this covariance estimator through influence functions. The influence function associated with L1 moment model and the L2 moment model are derived for i.i.d. setting, however, we have assessed that the influence function based covariance estimator is approximately unbiased when the clustering effect is small. We list several topics to address in future research:

- Additional simulation studies are needed in order to extend the Gini index model to multiple predictors. In addition, working out the influence function of the L2 moment model for clustered data remains a challenge.

- As is mentioned in Chapter 4, the proposed L2 moment model is merely an approximation of the expression (4.9) on page 116. Therefore, it is possible that in other settings, better approximations will have to considered. It is important to conduct research on a more general model approximation to the expression (4.9), by for example, replacing the exponential function by a double exponential function in model (4.10) or by simulating Poisson-distributed responses according to a log-linear model with quadratic covariates. In addition to studying the model adequacy, its impact on the interpretation should also be addressed.

- The applicability of the L2 moment model can be increased by introducing flexibility to the predictors, by, for example allowing additive models:

$$\mathrm{E}(|Y - Y'| \mid \boldsymbol{X}, \boldsymbol{X}') = g^{-1}[\beta_0 + \sum_{j=1}^{p} f_j(X_j, X_j')],$$

where X_j is the j^{th} entry of $\boldsymbol{X}$. The functions $f_j(X_j, X_j')$ are unknown smooth functions and they can be estimated from the data using algorithms such as backfitting.

- In this dissertation we have only demonstrated the method when information of species abundance is available, however the method can be extended for species incidence data. One biodiversity index used for absence/presence data is the taxonomic distinctness Δ^* **?**, it simply possesses the following form:

$$\Delta^* = \frac{\sum_i \sum_j w_{ij}}{n(n-1)},$$

where n is the total number of species observed and w_{ij} is the 'distance' between all pairs of species. This distance is often defined as the path lengths between successive taxonomic levels (species to genera, genera to families etc). For example, for a pair of two observed species, $w_{ij} = 1$ if they come from the same genus and $w_{ij} = 2$ if they are in the same family. Conventionally the increment is often set to be constant and most likely to be 1. The pseudo-outcome of the pairwise comparison is integers $0, 1, \ldots, m$, here m is often up to 5 or 6. The semiparametric Gini index model can then be expressed as:

$$\mathrm{E}(w \mid \boldsymbol{X}, \boldsymbol{X}') = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}),$$

the function $m(\cdot)$ must be symmetric and it is constrained to be related to a linear prediction $\boldsymbol{Z}^T\boldsymbol{\beta}$, then

$$m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}) = g^{-1}(\boldsymbol{Z}^T\boldsymbol{\beta}),$$

one possible choice of $g(\cdot)$ is *probit* function.

In Chapter 5 we have applied the methods developed in Chapter 2 and 4 to the human infant gut microbiome data. The applications demonstrated how to apply the method to human microbiome studies, which nowadays receive a lot of attention. We have accounted for the data normalization by using library size as the offset in the Poisson and ZINB-based approach. Interesting relationships between type 1 diabetes and various dietary intakes are revealed from the resulting ordination diagrams. Studies on the significance test of the effect of the diet is a possible starting point for further research. The association between microbiome diversity and age has been studied through the proposed semiparametric Gini index model in Chapter 5 as well. To obtain the covariance estimator, we suggested the use of a bootstrapping technique to account for the clustered structure. However, in the future, a solid theoretical foundation of the L2 moment model influence function needs to be built. Besides, the way to accommodate the normalisation of the microbiome data in the proposed L2 moment model is not yet resolved. Although it is common to use relative abundance (abundance divided by the library size), it is not applied in Chapter 5 since we have only assessed the form of $(X + X')$ under the assumption of Poisson distribution of the abundance Y and Y'. This lead to future research on the extension of the L2 moment model to adapt non-integer pseudo-outcomes.

Appendix A

Theory

A.1 Constrained Ordination Analysis in the Presence of Zero Inflation

A.1.1 The ZIP Distribution

A ZIP is defined as a mixture of a Poisson and a distribution degenerated at zeros, and it is given by

$$\Pr(Y_{ik} = y_{ik}) = \begin{cases} (1 - \pi_{ik}) \frac{e^{-\lambda_{ik}} \lambda_{ik}^{y_{ik}}}{y_{ik}!} & y_{ik} = 1, 2, \ldots \\ \pi_{ik} + (1 - \pi_{ik}) e^{-\lambda_{ik}} & y_{ik} = 0. \end{cases}$$

See **?** for more details.

A.1.2 The ZINB Distribution

A ZINB arises as a mixture of a Negative Binomial and a distribution degenerated at zeros, and it is given by

$$\Pr(Y_{ik} = y_{ik}) = \begin{cases} (1-\pi_{ik})\frac{\Gamma(y_{ik}+\rho ik)}{y!\Gamma(\rho_{ik})}(1+\lambda_{ik}/\rho_{ik})^{-\rho_{ik}}(1+\rho_{ik}/\lambda_{ik})^{-y_{ik}} & y_{ik} = 1, 2, \dots \\ \pi_{ik} + (1-\pi_{ik})(1+\lambda_{ik}/\rho_{ik})^{-\rho_{ik}} & y_{ik} = 0. \end{cases}$$

Note that when $\pi_{ik}\ to0$ and $\rho_{ik} \to \infty$, ZINB distribution collapses to ZIP distribution. We refer to **?** for more details.

A.1.3 The HZTP Distribution

A HZTP is composed by a Binomial distribution for the hurdle at zero and a zero-truncated Poisson, and it is given by

$$\Pr(Y_{ik} = y_{ik}) = \begin{cases} (1-\pi_{ik})\frac{\lambda_{ik}^{y_{ik}}}{y_{ik}!(e^{\lambda_{ik}}-1)} & y_{ik} = 1, 2, \dots \\ \pi_{ik} & y_{ik} = 0. \end{cases}$$

The HZTP is a special case of hurdle Poisson. See **?** for more details about the Poisson hurdle specification.

A.1.4 The HZTNB Distribution

HZTNB can be generalized to Hurdle Negative Binomial Model, which is widely considered as the most popular hurdle model in practice [**?**]. The probability mass function

of HZTNB is given by

$$\Pr(Y_{ik} = y_{ik}) = \begin{cases} (1-\pi_{ik})\frac{\Gamma(y_{ik}+\rho ik)(1+\rho_{ik}/\lambda_{ik})^{-y_{ik}}}{y!\Gamma(\rho_{ik})((1+\lambda_{ik}/\rho_{ik})^{\rho_{ik}}-1)} & y_{ik} = 1, 2, \ldots \\ \pi_{ik} & y_{ik} = 0. \end{cases}$$

A.2 Constrained Ordination Analysis with Enrichment of Bell-Shaped Response Functions

Iterative Reweighted Least Squares

Consider the first order Taylor expansion

$$\begin{aligned} f_k(z_i;\boldsymbol{\beta}_k) &= \exp(\boldsymbol{\beta}_k^T \boldsymbol{w}_i) \\ &\approx \exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i) + \left.\frac{\partial f_k(z_i;\boldsymbol{\beta}_k)}{\partial \boldsymbol{\beta}_k}\right|_{\tilde{\boldsymbol{\beta}}_k} (\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k) \\ &\approx \exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i) + \exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i)\boldsymbol{w}_i^T(\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k) \\ &\approx (1 - \boldsymbol{w}_i^T\tilde{\boldsymbol{\beta}}_k)\exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i) + \exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i)\boldsymbol{w}_i^T\boldsymbol{\beta}_k. \end{aligned}$$

Let $\boldsymbol{\lambda}_k^T = (f_k(z_1;\boldsymbol{\beta}_k), \ldots, f_k(z_n;\boldsymbol{\beta}_k))$ for which the Taylor expansion gives

$$\boldsymbol{\lambda}_k^T \approx \tilde{\boldsymbol{D}}_{1k}\tilde{\boldsymbol{\lambda}}_k + \tilde{\boldsymbol{D}}_{2k}\boldsymbol{W}\boldsymbol{\beta}_k,$$

with the matrices as defined in Section Bell-Shape Enriched Constrained Ordination analysis.

Upon introducing matrix notation in Equation 5 in Section 3.2.2 to replace the summation, and upon using the Taylor expansion, the estimating equation becomes approxi-

mately Equation 6 in Section Penalised Maximum Likelihood.

A.2.1 Newton-Raphson for the maximisation of LLR

The $\boldsymbol{\alpha}$ maximising Equitation 9 in Section 3.2.3 is found by solving the equation

$$0 = \boldsymbol{U}(\boldsymbol{Y}, \boldsymbol{\alpha}) = \frac{\partial \log \mathsf{LR}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = \sum_{i=1}^{n}\sum_{k=1}^{s} \frac{\partial \log p_k(y_{ik} \mid z_i, \boldsymbol{\beta}_k)}{\partial \boldsymbol{\alpha}} - \sum_{i=1}^{n}\sum_{k=1}^{s} \frac{\partial \log p(y_{ik} \mid z_i, \boldsymbol{\beta})}{\partial \boldsymbol{\alpha}}. \tag{A.1}$$

The first term equals

$$\sum_{i=1}^{n}\sum_{k=1}^{s}(y_{ik} - \exp(\boldsymbol{\beta}_k^T \boldsymbol{w}_i))\boldsymbol{A}_{ik}$$

with

$$\boldsymbol{A}_{ik} = \boldsymbol{\beta}_k^T \frac{\partial \boldsymbol{w}_i}{\partial z}\boldsymbol{x}_i.$$

Similarly, the second term of (A.1) can be written as

$$\sum_{i=1}^{n}\sum_{k=1}^{s}(y_{ik} - \exp(\boldsymbol{\beta}^T \boldsymbol{w}_i^*))\boldsymbol{A}_i$$

with $\boldsymbol{A}_i = \boldsymbol{\beta}^T \frac{\partial \boldsymbol{w}_i^*}{\partial z}\boldsymbol{x}_i$ and in which we use $\boldsymbol{w}_i^*$ to stress that the structure of the common model may be different from the structure of the species-specific models (however, in most applications $\boldsymbol{w}_i^* = \boldsymbol{w}_i$).

Newton-Raphson requires the Hessian matrix

$$\begin{aligned}
\boldsymbol{H}(\boldsymbol{Y}, \boldsymbol{\alpha}) &= \frac{\partial \boldsymbol{U}(\boldsymbol{Y}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \\
&= \sum_{i=1}^{n}\sum_{k=1}^{s}\Big[-\exp(\boldsymbol{\beta}_k^T \boldsymbol{w}_i)\boldsymbol{A}_{ik}\boldsymbol{A}_{ik}^T + (y_{ik} - \exp(\boldsymbol{\beta}_k^T \boldsymbol{w}_i))\boldsymbol{A}'_{ik} \\
&\quad + \exp(\boldsymbol{\beta}^T \boldsymbol{w}_i^*)\boldsymbol{A}_i^T \boldsymbol{A}_i - (y_{ik} - \exp(\boldsymbol{\beta}^T \boldsymbol{w}_i^*))\boldsymbol{A}'_i\Big],
\end{aligned}$$

where

$$\boldsymbol{A}'_{ik} = \frac{\partial \boldsymbol{A}_{ik}}{\partial \boldsymbol{\alpha}} = \boldsymbol{\beta}_k^T \frac{\partial^2 \boldsymbol{w}_i}{\partial z^2} \boldsymbol{x}_i \boldsymbol{x}_i^T \text{ and } \boldsymbol{A}'_i = \frac{\partial \boldsymbol{A}_i}{\partial \boldsymbol{\alpha}} = \boldsymbol{\beta}^T \frac{\partial^2 \boldsymbol{w}_i^*}{\partial z^2} \boldsymbol{x}_i \boldsymbol{x}_i^T.$$

Newton-Raphson starts with an initial $\boldsymbol{\alpha}$, say $\boldsymbol{\alpha}^{(0)}$, which is updated by

$$\boldsymbol{\alpha}^{(m+1)} = \boldsymbol{\alpha}^{(m)} - \boldsymbol{H}^{-1}(\boldsymbol{Y}, \boldsymbol{\alpha}^{(m)}) \boldsymbol{U}(\boldsymbol{Y}, \boldsymbol{\alpha}^{(m)}), \tag{A.2}$$

$m = 0, 1, \ldots$ (until convergence).

Fisher scoring also requires the expectation of the Hessian, which becomes, upon using $\mathrm{E}\,(Y_{ik} \mid z_i) = \exp(\boldsymbol{\beta}_k^T \boldsymbol{w}_i)$,

$$\begin{aligned} \boldsymbol{J}(\boldsymbol{\alpha}) &= \mathrm{E}\,(\boldsymbol{H}(\boldsymbol{Y}, \boldsymbol{\alpha}) \mid \boldsymbol{z}) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{s} \Big[-\exp(\boldsymbol{\beta}_k^T \boldsymbol{w}_i) \boldsymbol{A}_{ik} \boldsymbol{A}_{ik}^T \\ &\quad + \exp(\boldsymbol{\beta}^T \boldsymbol{w}_i^*) \boldsymbol{A}_i^T \boldsymbol{A}_i - (\exp(\boldsymbol{\beta}_k^T \boldsymbol{w}_i) - \exp(\boldsymbol{\beta}^T \boldsymbol{w}_i^*)) \boldsymbol{A}'_i \Big]. \end{aligned}$$

Newton-Raphson with fisher scoring replaces (A.2) with

$$\boldsymbol{\alpha}^{(m+1)} = \boldsymbol{\alpha}^{(m)} - \boldsymbol{J}^{-1}(\boldsymbol{\alpha}^{(m)}) \boldsymbol{U}(\boldsymbol{Y}, \boldsymbol{\alpha}^{(m)}).$$

A.2.2 Joint model fit

Table A.1: Comparison of the joint model fits for the first and second dimension from three ordination methods applied to the Antarctic lakes data. MSE gives the mean squared error calculated only among Bell-shaped species, MSE*stands for the mean squared error calculated from all species.

	BECOA	FCOA	CCA
$\sqrt{\text{MSE}}$	101.80	229.18	157.74
$\sqrt{\text{MSE}^*}$	152.68	152.98	157.74

A.2.3 Cross validation

In order to reduce the computational load, we propose a simplified cross-validation procedure in which the environmental gradient calculated from the complete data set, say $\hat{\boldsymbol{\alpha}}$, is kept constant throughout the calculations. In particular, the data set is randomly split into 10 equally large parts. We denote these subsets as $\boldsymbol{X}_j$ ($j = 1, \ldots, 10$). The procedure now iterates over the following steps: in the jth step, the jth subset $\boldsymbol{X}_j$ is removed and the remaining 9 subsets, denoted by $\boldsymbol{X}_{(-j)}$, serve as the training data, from which the environmental scores $\boldsymbol{z}_{\text{train}} = \boldsymbol{X}_{(-j)}\hat{\boldsymbol{\alpha}}$ are computed, as well as the β parameter estimates from the penalized regression procedure, say $\hat{\boldsymbol{\beta}}_{-j}$. For the testing data the scores $\boldsymbol{z}_{\text{test}} = \boldsymbol{X}_{(j)}\hat{\boldsymbol{\alpha}}$ and, upon using $\hat{\boldsymbol{\beta}}_{-j}$, the predictions of the abundances are computed. The latter, and the observed abundances in the test data set are subsequently used for the calculation of the residuals. Finally, after completion of the 10 cross validation cycles, all residuals are squared and summed, resulting in the cross-validation estimate SSE. This procedure is repeated for a sequence of δs.

The cross-validated average LLR can be obtained analogously.

A.2.4 Algorithm of absence/presence data

When the abundance data are replaced by the species absence ($Y_{ik} = 0$) / presence ($Y_{ik} > 0$) information, the probability of the presence of a species at certain location can be linked to environmental scores using the binomial distribution, and, for example, the logit link. In particular,

$$\pi_{ik} = \mathsf{P}\{Y_{ik} > 0 | z_i\} = \operatorname{expit}\left(a_k - \frac{(z_i - \mu_k)^2}{2t_k^2}\right).$$

The probability mass function is then written as

$$p_k(y_{ik} \mid z_i, \boldsymbol{\beta}_k) = \pi_{ik}^{\mathsf{I}[y_{ik}>0]}(1-\pi_{ik})^{\mathsf{I}[y_{ik}=0]}.$$

For this model, the score equation for the penalised maximum likelihood estimator of parameter β_{jk} becomes,

$$\sum_{i=1}^{n} \frac{\partial \log p_k(\boldsymbol{\beta}_k \mid y_{ik}, z_i)}{\partial \beta_{jk}} = \sum_{i=1}^{n} \left(y_{ik} - \frac{\exp(\boldsymbol{\beta}_k^T \boldsymbol{w}_i)}{1+\exp(\boldsymbol{\beta}_k^T \boldsymbol{w}_i)} \right) w_{ij} + n \frac{\partial \log g(\boldsymbol{\beta}_k)}{\partial \beta_{jk}} = 0. \quad \text{(A.3)}$$

Consider the first order Taylor expansion

$$\begin{aligned}
\operatorname{expit}(\boldsymbol{\beta}_k^T \boldsymbol{w}_i) &= \frac{\exp(\boldsymbol{\beta}_k^T \boldsymbol{w}_i)}{1+\exp(\boldsymbol{\beta}_k^T \boldsymbol{w}_i)} \\
&\approx \frac{\exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i)}{1+\exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i)} + \frac{\exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i)(1+\exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i))\boldsymbol{w}^{\boldsymbol{T}}{}_i - \exp^2(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i)\boldsymbol{w}_i^T}{(1+\exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i))^2}(\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k) \\
&\approx \frac{\exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i)}{1+\exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i)} + \left[\frac{\exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i)}{1+\exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i)} - \left(\frac{\exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i)}{1+\exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i)} \right)^2 \right] \boldsymbol{w}_i^T \boldsymbol{\beta}_k \\
&\quad + \left[\left(\frac{\exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i)}{1+\exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i)} \right)^2 - \frac{\exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i)}{1+\exp(\tilde{\boldsymbol{\beta}}_k^T \boldsymbol{w}_i)} \right] \boldsymbol{w}_i^T \tilde{\boldsymbol{\beta}}_k.
\end{aligned}$$

Let $\boldsymbol{\theta}_k^T = (\operatorname{expit}(\boldsymbol{\beta}_k^T \boldsymbol{w}_1), \ldots, \operatorname{expit}(\boldsymbol{\beta}_k^T \boldsymbol{w}_n))$ for which this Taylor expension gives

$$\boldsymbol{\theta}_k^T \approx \tilde{\boldsymbol{\theta}}_k + (\tilde{\boldsymbol{D}}_3 - \tilde{\boldsymbol{D}}_3^2)\boldsymbol{W}\boldsymbol{\beta}_k - (\tilde{\boldsymbol{D}}_3 - \tilde{\boldsymbol{D}}_3^2)\boldsymbol{W}\tilde{\boldsymbol{\beta}}_k,$$

where

$$\tilde{\boldsymbol{\theta}}_k = \operatorname{expit}(\boldsymbol{W}\tilde{\boldsymbol{\beta}}_k) \quad \tilde{\boldsymbol{D}}_3 = \mathsf{Diag}(\operatorname{expit}(\boldsymbol{\beta}_k^T \boldsymbol{w}_i))$$

and $\boldsymbol{W}$ is defined as in the paper. Given a $\tilde{\beta}_k$, Equation (A.3) gives a closed form for an updated estimate,

$$\hat{\beta}_k = \left(\boldsymbol{W}^T(\tilde{\boldsymbol{D}}_3 - \tilde{\boldsymbol{D}}_3^2)\boldsymbol{W} + n\gamma\boldsymbol{D}^{-1}\right)^{-1}\left[(\boldsymbol{Y} - \tilde{\boldsymbol{\theta}}_k)^T\boldsymbol{W} + \tilde{\beta}_k^T\boldsymbol{W}^T(\tilde{\boldsymbol{D}}_3 - \tilde{\boldsymbol{D}}_3^2)\boldsymbol{W} + n\gamma\boldsymbol{\delta}^T\boldsymbol{D}^{-1}\right],$$

in which the terms $n\gamma\boldsymbol{D}^{-1}$ and $n\gamma\boldsymbol{\delta}^T\boldsymbol{D}^{-1}$ arise from the penalisation (see also algorithm 1 in Section Penalised Maximum Likelihood).

A.3 Semiparametric Gini Index Model

A.3.1 L-moments

The research on the theory of L-moment has a relatively short history as compared to that of the conventional moments [**??**]. The scattered results were gathered by **?** giving rise to the definition of L-moment as the linear combination of the expected order statistics of a population.

We first give the definition of the L-moment.

Definition 3 (L-moments)**.** *Let X be a real-valued random variable, $F(x)$ its cumulative distribution function and $x(F)$ the quantile function. The ordered statistics of size n drawn from the distribution of X are denoted as $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$. The r^{th} L-moment of X, λ_r, is then defined as*

$$\lambda_r \equiv r^{-1}\sum_{k=0}^{r-1}\begin{pmatrix} r-1 \\ k \end{pmatrix}\mathrm{E}(X_{r-k:r}).$$

Analogously to the conventional moments, L-moments can be employed to describe the probability distribution. The first L-moment, λ_1, is defined as:

$$\lambda_1 \int_0^1 x(F)dF = \mathrm{E}(X) \tag{A.4}$$

The λ_1 is a location parameter and is often referred to as *L-location*, and it is just the expectation of X. The first L-moment is thus equal to the first moment.

The second L-moment, λ_2, is a measure of variability and has the expression:

$$\lambda_2 = \int_0^1 x(F)(2F-1)dF = \frac{1}{2}\mathrm{E}(X_{2:2} - X_{1:2}), \tag{A.5}$$

The second L-moment is also called L-scale. Since $\lambda_2 = \mathrm{E}(|X_1 - X_2|)/2$, the second L-moment can be estimated as

$$\hat{\lambda}_2 = \frac{n(n-1)}{2} \frac{\sum_{1 \le i < j \le n} |X_i - X_j|}{2}.$$

The second L-moment, as a measure of variability, has the advantage of being less sensitive to outliers than the ordinary second moment [**?**].

The third and fourth L-moments are named L-skewness and L-kurtosis respectively and in this thesis, only the first two L-moments is of primary interest. For notational simplicity, from now on we will use *L1 moment* to refer to the first L-moment (thus the expectation) and *L2 moment* for the second L-moment.

A.3.2 Relationship between Gini index and L-moments

For a random variable X with cumulative distribution function $F(x)$, the following definition of Lorenz curve is given by **?**:

$$L(u) = \mu^{-1} \int_0^u F^{-1}(x)dx = \mu^{-1} \int_0^u x(F)dF, \quad 0 \leq u \leq 1,$$

where μ is the expectation of X. In the rest of this section, we show how the Gini index is related to the first two L-moments. Recall that the Gini index is the area between Lorenz curve and the line of equality, A, and B is the area under the Loren curve (see Figure 1.6). As $A + B = 0.5$, another way to express the Gini index is $(1 - 2B)/2$, where

$$B = \int_0^1 L(u)du = \int_0^1 \mu^{-1} \int_0^u x(F)dFdu.$$

By reversing the order of integration we have

$$\begin{aligned} B = \mu^{-1} \int_0^1 \int_0^u x(F)dFdu &= \mu^{-1} \int_0^1 \int_F^1 x(F)dudF \\ &= \mu^{-1} \int_0^1 [x(F) - x(F)F]dF \\ &= \mu^{-1} \left[\int_0^1 x(F)dF - \int_0^1 x(F)FdF \right]. \end{aligned}$$

From Equation (A.5) we have

$$\lambda_2 = 2 \int_0^1 x(F)FdF - \int_0^1 x(F)dF = 2 \int_0^1 x(F)FdF - \mu.$$

The last equality follows from Equation (A.4). If G denotes the Gini index, it follows that

$$\begin{aligned} G &= \frac{1-2B}{2} \\ &= \frac{1-2\mu^{-1}\left[\int_0^1 x(F)dF - \int_0^1 x(F)FdF\right]}{2} \\ &= \frac{1-2\mu^{-1}\left[\mu - \int_0^1 x(F)FdF\right]}{2} \\ &= \frac{2\mu^{-1}\int_0^1 x(F)FdF - 1}{2} \\ &= \frac{2\int_0^1 x(F)FdF - \mu}{2\mu} \\ &= \frac{\lambda_2}{2\mu}. \end{aligned}$$

The Gini index is therefore equal to half of the ratio of L2 moment to L1 moment. This term is also named *L-CV* with *L* for L-moments and *CV* for coefficient of variation [**?**].

A.3.3 Estimate of the variance of $\hat{\boldsymbol{\alpha}}$

The variance of $\hat{\boldsymbol{\alpha}}$ can be approximated by

$$\begin{aligned} n\text{Var}(\hat{\boldsymbol{\alpha}}) &\approx \text{Var}\Big((\text{E}\left(\boldsymbol{x}^T\exp(\boldsymbol{x}\boldsymbol{\alpha}_0)\boldsymbol{x}\right))^{-1}\boldsymbol{x}_i^T\left[y_i - \exp(\boldsymbol{x}_i\boldsymbol{\alpha}_0)\right]\Big) \\ &= (\text{E}\left(\boldsymbol{x}^T\exp(\boldsymbol{x}\boldsymbol{\alpha}_0)\boldsymbol{x}\right))^{-1} \\ &\quad \text{E}\left(\boldsymbol{x}_i^T\left[y_i - \exp(\boldsymbol{x}_i\boldsymbol{\alpha}_0)\right]^2\boldsymbol{x}_i\right) \\ &\quad (\text{E}\left(\boldsymbol{x}^T\exp(\boldsymbol{x}\boldsymbol{\alpha}_0)\boldsymbol{x}\right))^{-1^T} \end{aligned}$$

Set $\boldsymbol{\lambda}_1 = \text{diag}(\exp(\boldsymbol{X}\boldsymbol{\alpha}_0))$ and $\boldsymbol{\lambda}_2 = \text{diag}(\boldsymbol{Y} - \exp(\boldsymbol{X}\boldsymbol{\alpha}_0))$, the estimate of $\text{Var}(\hat{\boldsymbol{\alpha}})$ can be expressed in matrix form as follows

$$n\mathsf{Var}(\hat{\boldsymbol{\alpha}}) \approx \mathrm{E}\left(\boldsymbol{X}^T\boldsymbol{\lambda}_1\boldsymbol{X}\right)^{-1}\mathrm{E}\left(\boldsymbol{X}^T\boldsymbol{\lambda}_2\boldsymbol{\lambda}_2^T\boldsymbol{X}\right)\mathrm{E}\left(\boldsymbol{X}^T\boldsymbol{\lambda}_1\boldsymbol{X}\right)^{-1^T} \tag{A.6}$$

Estimate of the variance of $\hat{\beta}$

Set $\boldsymbol{B}(\boldsymbol{x}_i, y_i, \boldsymbol{\beta}_0) = \mathrm{E}\left((\boldsymbol{x}_i + \boldsymbol{x}_j)^T \exp((\boldsymbol{x}_i + \boldsymbol{x}_j)\boldsymbol{\beta}_0)\Big[|y_i - y_j| - \exp((\boldsymbol{x}_i + \boldsymbol{x}_j)\boldsymbol{\beta}_0)\Big] \mid \boldsymbol{x}_i, y_i\right)$ the variance of $\hat{\beta}$can be expressed as

$$\begin{aligned} n\mathsf{VAR}(\hat{\boldsymbol{\beta}}) &\approx \mathsf{VAR}\Big(\boldsymbol{C}\boldsymbol{B}(\boldsymbol{x}_i, y_i, \boldsymbol{\beta}_0)\Big) \\ &= \boldsymbol{C}\mathrm{E}\left(\boldsymbol{B}(\boldsymbol{x}_i, y_i, \boldsymbol{\beta}_0)\boldsymbol{B}^T(\boldsymbol{x}_i, y_i, \boldsymbol{\beta}_0)\right)\boldsymbol{C}^T \end{aligned}$$

A.3.4 Estimate of the covariance of $\hat{\alpha}$ and $\hat{\beta}$

$$\begin{aligned} n\mathsf{Cov}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) &\approx \mathsf{COV}\Big((\mathrm{E}\left(\boldsymbol{x}^T\exp(\boldsymbol{x}\boldsymbol{\alpha}_0)\boldsymbol{x}\right))^{-1}\boldsymbol{x}_i^T\Big[y_i - \exp(\boldsymbol{x}_i\boldsymbol{\alpha}_0)\Big], \boldsymbol{C}\boldsymbol{B}(\boldsymbol{x}_i, y_i)\Big) \\ &= (\mathrm{E}\left(\boldsymbol{x}^T\exp(\boldsymbol{x}\boldsymbol{\alpha}_0)\boldsymbol{x}\right))^{-1}\mathrm{E}\left(\boldsymbol{x}_i^T\Big[y_i - \exp(\boldsymbol{x}_i\boldsymbol{\alpha}_0)\Big]\boldsymbol{B}^T(\boldsymbol{x}_i, y_i, \boldsymbol{\beta}_0)\right)\boldsymbol{C}^T, \end{aligned}$$

Appendix B

Data set

B.1 Antarctic lakes data sets

	Min	Q1	Median	Mean	Q3	Max.
Sandaracinobacter	0.0000	0.0000	0.0000	2.7110	0.0000	40.0000
Smithella	0.0000	0.0000	0.0000	7.5330	0.0000	221.0000
Sphingobium	0.0000	0.0000	0.0000	0.7111	0.0000	15.0000
unclassified_Acidimicrobiaceae	0.0000	0.0000	0.0000	0.5778	0.0000	12.0000
unclassified_Nocardiaceae	0.0000	0.0000	0.0000	0.4222	0.0000	7.0000
Acetobacterium	0.0000	0.0000	0.0000	0.9556	0.0000	16.0000
Aequorivita	0.0000	0.0000	0.0000	5.8890	0.0000	96.0000
Alkalibacterium	0.0000	0.0000	0.0000	152.7000	0.0000	4463.0000
Aminobacter	0.0000	0.0000	0.0000	1.8890	0.0000	43.0000
Bdellovibrio	0.0000	0.0000	0.0000	0.6444	0.0000	7.0000
Desulfovibrio	0.0000	0.0000	0.0000	2.7560	0.0000	48.0000

Gaetbulibacter	0.0000	0.0000	0.0000	0.6000	0.0000	8.0000
Gp2	0.0000	0.0000	0.0000	0.6667	0.0000	6.0000
Halomonas	0.0000	0.0000	0.0000	4.7560	0.0000	52.0000
Oxobacter	0.0000	0.0000	0.0000	1.6440	0.0000	42.0000
Phaeobacter	0.0000	0.0000	0.0000	0.7556	0.0000	11.0000
Pseudonocardia	0.0000	0.0000	0.0000	2.1780	0.0000	37.0000
Psychroflexus	0.0000	0.0000	0.0000	85.3800	0.0000	1578.0000
Schlesneria	0.0000	0.0000	0.0000	2.5110	0.0000	44.0000
Seohaeicola	0.0000	0.0000	0.0000	8.9560	0.0000	244.0000
Silanimonas	0.0000	0.0000	0.0000	0.7111	0.0000	8.0000
Staphylococcus	0.0000	0.0000	0.0000	6.2440	0.0000	239.0000
Tatlockia	0.0000	0.0000	0.0000	0.9111	0.0000	12.0000
unclassified_Coriobacteriaceae	0.0000	0.0000	0.0000	4.3110	0.0000	107.0000
unclassified_Cystobacteraceae	0.0000	0.0000	0.0000	0.6889	0.0000	11.0000
unclassified_Kineosporiaceae	0.0000	0.0000	0.0000	2.2890	0.0000	57.0000
unclassified_Nakamurellaceae	0.0000	0.0000	0.0000	4.1330	0.0000	49.0000
Altererythrobacter	0.0000	0.0000	0.0000	2.0220	0.0000	22.0000
Arcicella	0.0000	0.0000	0.0000	7.8000	0.0000	208.0000
Bellilinea	0.0000	0.0000	0.0000	3.4440	0.0000	51.0000
Cellulomonas	0.0000	0.0000	0.0000	4.2670	0.0000	81.0000
Chryseobacterium	0.0000	0.0000	0.0000	2.6000	0.0000	35.0000
Cryomorpha	0.0000	0.0000	0.0000	1.0670	0.0000	10.0000
Defluviicoccus	0.0000	0.0000	0.0000	8.5110	0.0000	220.0000
Deinococcus	0.0000	0.0000	0.0000	7.5110	0.0000	127.0000

GpV	0.0000	0.0000	0.0000	6.1780	0.0000	190.0000
GpXI	0.0000	0.0000	0.0000	1.4000	0.0000	19.0000
Lewinella	0.0000	0.0000	0.0000	3.5330	0.0000	73.0000
Marinobacter	0.0000	0.0000	0.0000	11.8900	0.0000	301.0000
Persicivirga	0.0000	0.0000	0.0000	34.2700	0.0000	1401.0000
Psychrobacter	0.0000	0.0000	0.0000	55.0000	0.0000	1266.0000
Rhodococcus	0.0000	0.0000	0.0000	1.8670	0.0000	56.0000
Spirochaeta	0.0000	0.0000	0.0000	1.5330	0.0000	23.0000
Stenotrophomonas	0.0000	0.0000	0.0000	3.1330	0.0000	72.0000
Tessaracoccus	0.0000	0.0000	0.0000	4.1330	0.0000	43.0000
Verrucomicrobium	0.0000	0.0000	0.0000	1.0890	0.0000	11.0000
Blastopirellula	0.0000	0.0000	0.0000	2.1330	0.0000	33.0000
Desulforhopalus	0.0000	0.0000	0.0000	10.0200	0.0000	221.0000
Gracilimonas	0.0000	0.0000	0.0000	67.4400	0.0000	1168.0000
Haematobacter	0.0000	0.0000	0.0000	6.3110	0.0000	84.0000
Planococcus	0.0000	0.0000	0.0000	22.9300	0.0000	608.0000
Rubritepida	0.0000	0.0000	0.0000	8.8670	0.0000	195.0000
Rudaea	0.0000	0.0000	0.0000	3.1560	0.0000	36.0000
Sporotalea	0.0000	0.0000	0.0000	2.8220	0.0000	35.0000
Streptomyces	0.0000	0.0000	0.0000	1.4440	0.0000	27.0000
Tissierella	0.0000	0.0000	0.0000	27.4900	0.0000	755.0000
unclassified_Nocardioidaceae	0.0000	0.0000	0.0000	1.4890	0.0000	25.0000
Cyclobacterium	0.0000	0.0000	0.0000	4.2000	0.0000	104.0000
Dokdonella	0.0000	0.0000	0.0000	3.8000	0.0000	51.0000

Kofleria	0.0000	0.0000	0.0000	1.1330	0.0000	12.0000
Prolixibacter	0.0000	0.0000	0.0000	2.6440	0.0000	36.0000
Pseudoalteromonas	0.0000	0.0000	0.0000	1.0000	0.0000	17.0000
Sporichthya	0.0000	0.0000	0.0000	1.8670	0.0000	20.0000
Steroidobacter	0.0000	0.0000	0.0000	2.8890	0.0000	57.0000
Subsaxibacter	0.0000	0.0000	0.0000	14.5100	0.0000	161.0000
Thermomonas	0.0000	0.0000	0.0000	6.7560	0.0000	102.0000
unclassified_Polyangiaceae	0.0000	0.0000	0.0000	1.6890	0.0000	25.0000
Arenimonas	0.0000	0.0000	0.0000	4.8890	0.0000	110.0000
Carnobacterium	0.0000	0.0000	0.0000	255.4000	0.0000	3884.0000
Gelidibacter	0.0000	0.0000	0.0000	37.5300	0.0000	752.0000
Geminicoccus	0.0000	0.0000	0.0000	4.1330	0.0000	52.0000
Kineococcus	0.0000	0.0000	0.0000	5.5110	0.0000	89.0000
Microbacterium	0.0000	0.0000	0.0000	1.9330	0.0000	48.0000
Nannocystis	0.0000	0.0000	0.0000	5.4220	0.0000	90.0000
Nitrospira	0.0000	0.0000	0.0000	4.4890	0.0000	77.0000
Ochrobactrum	0.0000	0.0000	0.0000	2.1560	0.0000	30.0000
Pseudoxanthomonas	0.0000	0.0000	0.0000	5.8220	0.0000	110.0000
Rubellimicrobium	0.0000	0.0000	0.0000	7.2220	0.0000	70.0000
Rubrobacter	0.0000	0.0000	0.0000	8.6670	0.0000	303.0000
Rudanella	0.0000	0.0000	0.0000	3.0000	0.0000	24.0000
Streptococcus	0.0000	0.0000	0.0000	5.7780	0.0000	199.0000
Aquabacterium	0.0000	0.0000	0.0000	4.1560	3.0000	40.0000
Burkholderia	0.0000	0.0000	0.0000	4.4890	2.0000	60.0000

Desulfocapsa	0.0000	0.0000	0.0000	4.2000	2.0000	91.0000
Gp1	0.0000	0.0000	0.0000	7.0890	2.0000	173.0000
Hoeflea	0.0000	0.0000	0.0000	8.9110	3.0000	179.0000
Marmoricola	0.0000	0.0000	0.0000	7.1780	2.0000	85.0000
Propionivibrio	0.0000	0.0000	0.0000	6.2220	3.0000	121.0000
Pseudorhodobacter	0.0000	0.0000	0.0000	15.0400	2.0000	268.0000
Rhodoplanes	0.0000	0.0000	0.0000	1.9110	3.0000	19.0000
Terrimonas	0.0000	0.0000	0.0000	2.6890	2.0000	52.0000
Thiobacillus	0.0000	0.0000	0.0000	3.3560	2.0000	67.0000
Variovorax	0.0000	0.0000	0.0000	5.4440	5.0000	52.0000
Winogradskyella	0.0000	0.0000	0.0000	4.3110	2.0000	37.0000
Anaerovorax	0.0000	0.0000	0.0000	9.2220	2.0000	288.0000
Chelatococcus	0.0000	0.0000	0.0000	3.7780	2.0000	49.0000
Cryobacterium	0.0000	0.0000	0.0000	19.7800	3.0000	425.0000
Gp7	0.0000	0.0000	0.0000	6.9780	2.0000	117.0000
Humicoccus	0.0000	0.0000	0.0000	9.9330	5.0000	160.0000
Jannaschia	0.0000	0.0000	0.0000	4.6000	2.0000	47.0000
Microcella	0.0000	0.0000	0.0000	4.0220	4.0000	38.0000
Paenibacillus	0.0000	0.0000	0.0000	9.6890	2.0000	251.0000
Persicitalea	0.0000	0.0000	0.0000	40.0000	4.0000	743.0000
Filomicrobium	0.0000	0.0000	0.0000	5.3560	4.0000	112.0000
Herbaspirillum	0.0000	0.0000	0.0000	7.8670	4.0000	108.0000
Ideonella	0.0000	0.0000	0.0000	5.5110	4.0000	57.0000
Lactobacillus	0.0000	0.0000	0.0000	3.9780	3.0000	51.0000

Lysobacter	0.0000	0.0000	0.0000	11.3600	7.0000	186.0000
Maribacter	0.0000	0.0000	0.0000	18.9600	9.0000	456.0000
Methylocystis	0.0000	0.0000	0.0000	3.5110	3.0000	68.0000
Dehalogenimonas	0.0000	0.0000	0.0000	18.5800	4.0000	382.0000
Nitriliruptor	0.0000	0.0000	0.0000	22.2900	7.0000	386.0000
Pseudolabrys	0.0000	0.0000	0.0000	7.8440	5.0000	109.0000
Bacillus	0.0000	0.0000	0.0000	6.2000	7.0000	72.0000
Demequina	0.0000	0.0000	0.0000	21.6400	9.0000	241.0000
Desulfosporosinus	0.0000	0.0000	0.0000	6.6220	6.0000	69.0000
Nocardioides	0.0000	0.0000	0.0000	16.2000	6.0000	215.0000
Oceanicola	0.0000	0.0000	0.0000	5.6670	8.0000	47.0000
Opitutus	0.0000	0.0000	0.0000	3.7330	4.0000	45.0000
unclassified_Intrasporangiaceae	0.0000	0.0000	0.0000	10.0900	5.0000	99.0000
Bosea	0.0000	0.0000	0.0000	6.9330	5.0000	94.0000
Caulobacter	0.0000	0.0000	0.0000	7.6220	4.0000	145.0000
Sulfitobacter	0.0000	0.0000	0.0000	80.2200	27.0000	1203.0000
Corynebacterium	0.0000	0.0000	0.0000	14.1600	6.0000	473.0000
Gillisia	0.0000	0.0000	0.0000	99.1600	27.0000	2165.0000
Pedobacter	0.0000	0.0000	0.0000	13.1300	9.0000	230.0000
Propionibacterium	0.0000	0.0000	0.0000	23.8000	5.0000	846.0000
Pseudomonas	0.0000	0.0000	0.0000	12.6900	3.0000	355.0000
Sandarakinorhabdus	0.0000	0.0000	0.0000	3.1560	4.0000	32.0000
unclassified_Propionibacteriaceae	0.0000	0.0000	0.0000	67.8400	29.0000	1330.0000
Geobacter	0.0000	0.0000	0.0000	7.2670	5.0000	151.0000

GpIIa	0.0000	0.0000	0.0000	67.2000	11.0000	822.0000
Hymenobacter	0.0000	0.0000	0.0000	21.8000	13.0000	270.0000
Phenylobacterium	0.0000	0.0000	0.0000	8.9560	9.0000	103.0000
Rhizobium	0.0000	0.0000	0.0000	5.8000	5.0000	88.0000
Roseovarius	0.0000	0.0000	0.0000	184.0000	17.0000	4883.0000
Truepera	0.0000	0.0000	0.0000	24.7100	19.0000	250.0000
Amaricoccus	0.0000	0.0000	0.0000	16.0400	11.0000	264.0000
unclassified_Clostridiaceae_1	0.0000	0.0000	0.0000	18.9600	9.0000	259.0000
Algoriphagus	0.0000	0.0000	0.0000	28.9600	22.0000	365.0000
Aquicella	0.0000	0.0000	0.0000	36.8700	5.0000	843.0000
Hydrogenophaga	0.0000	0.0000	0.0000	17.2000	26.0000	168.0000
Gp3	0.0000	0.0000	0.0000	24.9300	26.0000	265.0000
Mycobacterium	0.0000	0.0000	0.0000	19.0700	11.0000	285.0000
Paracoccus	0.0000	0.0000	0.0000	10.4700	6.0000	125.0000
Rhodovarius	0.0000	0.0000	0.0000	46.2900	32.0000	880.0000
Spirosoma	0.0000	0.0000	0.0000	48.8400	33.0000	774.0000
Acetivibrio	0.0000	0.0000	2.0000	45.5600	36.0000	551.0000
Arthrobacter	0.0000	0.0000	2.0000	22.3600	9.0000	552.0000
Bradyrhizobium	0.0000	0.0000	2.0000	7.5780	9.0000	79.0000
Hyphomicrobium	0.0000	0.0000	2.0000	15.2200	9.0000	120.0000
Erythrobacter	0.0000	0.0000	2.0000	5.7780	7.0000	60.0000
Roseococcus	0.0000	0.0000	3.0000	22.7300	28.0000	163.0000
Chloroflexus	0.0000	0.0000	6.0000	24.7100	25.0000	151.0000
Devosia	0.0000	0.0000	5.0000	26.0900	28.0000	235.0000

Loktanella	0.0000	0.0000	4.0000	108.8000	111.0000	929.0000
Prosthecomicrobium	0.0000	0.0000	4.0000	29.4900	30.0000	245.0000
Conexibacter	0.0000	0.0000	3.0000	11.5800	12.0000	179.0000
GpXIII	0.0000	0.0000	16.0000	116.7000	64.0000	1352.0000
Massilia	0.0000	0.0000	3.0000	21.3100	9.0000	225.0000
Ferruginibacter	0.0000	0.0000	6.0000	22.1600	23.0000	206.0000
Gp16	0.0000	0.0000	4.0000	50.2400	24.0000	1440.0000
Gp4	0.0000	0.0000	11.0000	43.1800	34.0000	426.0000
Mesorhizobium	0.0000	0.0000	4.0000	13.0700	17.0000	103.0000
Pirellula	0.0000	0.0000	5.0000	22.8000	25.0000	304.0000
Porphyrobacter	0.0000	0.0000	9.0000	140.4000	113.0000	1772.0000
Acinetobacter	0.0000	0.0000	7.0000	67.4700	38.0000	840.0000
Gp6	0.0000	0.0000	5.0000	33.6700	22.0000	488.0000
Polaromonas	0.0000	0.0000	6.0000	44.8900	25.0000	444.0000
Methylobacterium	0.0000	0.0000	5.0000	22.5800	13.0000	350.0000
Rhodoferax	0.0000	0.0000	6.0000	45.2900	21.0000	1234.0000
Zavarzinella	0.0000	0.0000	14.0000	38.0900	51.0000	262.0000
Clostridium	0.0000	0.0000	11.0000	76.1100	59.0000	679.0000
GpVI	0.0000	0.0000	30.0000	229.5000	469.0000	1716.0000
Luteolibacter	0.0000	0.0000	5.0000	18.4900	24.0000	133.0000
Novosphingobium	0.0000	0.0000	10.0000	16.6000	25.0000	96.0000
Methylibium	0.0000	0.0000	10.0000	27.3100	28.0000	268.0000
Singulisphaera	0.0000	0.0000	9.0000	24.9800	44.0000	116.0000
Sphingopyxis	0.0000	0.0000	7.0000	48.3800	41.0000	908.0000

unclassified_Microbacteriaceae	0.0000	0.0000	7.0000	34.1300	30.0000	240.0000
Gemmata	0.0000	0.0000	27.0000	110.8000	114.0000	815.0000
Leifsonia	0.0000	0.0000	9.0000	29.9600	40.0000	167.0000
Rhodopirellula	0.0000	0.0000	10.0000	25.1600	38.0000	164.0000
Sphaerobacter	0.0000	0.0000	10.0000	29.8000	38.0000	146.0000
Planctomyces	0.0000	0.0000	8.0000	18.8900	22.0000	141.0000
Roseomonas	0.0000	0.0000	25.0000	83.4700	100.0000	706.0000
GpIV	0.0000	4.0000	19.0000	167.4000	139.0000	1994.0000
Flavobacterium	0.0000	3.0000	9.0000	36.2700	51.0000	238.0000
Brevundimonas	0.0000	11.0000	39.0000	209.0000	124.0000	6021.0000
GpI	0.0000	6.0000	61.0000	377.6000	295.0000	4041.0000
Iamia	0.0000	5.0000	13.0000	36.4000	48.0000	189.0000
Legionella	0.0000	10.0000	45.0000	84.2000	123.0000	413.0000
Sphingomonas	0.0000	7.0000	50.0000	85.4000	109.0000	631.0000
Ilumatobacter	0.0000	7.0000	28.0000	47.5300	65.0000	388.0000
Rhodobacter	0.0000	13.0000	48.0000	154.6000	223.0000	858.0000
Gemmatimonas	0.0000	10.0000	39.0000	120.9000	186.0000	735.0000
Haliscomenobacter	0.0000	7.0000	22.0000	56.6900	92.0000	320.0000
Caldilinea	0.0000	30.0000	95.0000	274.8000	297.0000	3395.0000
Ralstonia	0.0000	11.0000	21.0000	188.9000	81.0000	1831.0000

Table B.1: A summary of the observed abundance of the microbial species from the Antarctic lakes limnology study. Min represents the minimum, Q1 represents the first quantile, Median is the second quantile, Mean is the average, Q3 represents the third quantile and Max represents the maximum,

	Type	Mean	Standard deviation
depth	continuous	4.4311	7.5428
Conductivity	continuous	14.5895	32.5262
pH	continuous	7.6682	0.8266
TOC	continuous	10.5931	40.6015
DOC	continuous	10.2229	38.9278
Na	continuous	6339.6427	14387.3056
K	continuous	193.1465	426.8830
Ca	continuous	130.4607	272.0100
Mg	continuous	893.7431	2162.3954
Cl	continuous	12590.4907	27966.3146
SO4	continuous	728.2984	1715.3027
NH4-N	continuous	0.5788	2.5137
Silicate-Si	continuous	1.7270	2.2854

Table B.2: A summary of the environmental data set.

B.2 The Dutch dune spider data set

	Min	Q1	Median	Mean	Q3	Max
Alopacce	0.0000	0.0000	2.0000	6.2140	12.0000	29.0000
Alopcune	0.0000	0.0000	1.0000	5.3930	6.2500	43.0000
Alopfabr	0.0000	0.0000	0.0000	3.4640	3.0000	20.0000
Arctlute	0.0000	0.0000	0.0000	0.9286	0.2500	12.0000
Arctperi	0.0000	0.0000	0.0000	1.3930	0.0000	18.0000
Auloalbi	0.0000	0.0000	0.0000	4.6430	6.2500	30.0000
Pardlugu	0.0000	0.0000	1.0000	4.5360	3.5000	55.0000
Pardmont	0.0000	0.7500	4.5000	16.0400	22.5000	96.0000
Pardnigr	0.0000	0.0000	1.0000	14.5000	15.0000	135.0000
Pardpull	0.0000	0.0000	0.5000	20.7900	39.0000	105.0000
Trocterr	0.0000	2.0000	22.5000	34.6800	63.5000	118.0000
Zoraspin	0.0000	0.0000	2.0000	6.6070	6.7500	34.0000

Table B.3: A summary of the observed abundance of 12 types of spiders from the Dutch dune area. Min represents the minimum, Q1 represents the first quantile, Median is the second quantile, Mean is the average, Q3 represents the third quantile and Max represents the maximum,

	Type	Mean	Standard deviation
WaterCon	continuous	2.4713	0.8088
BareSand	continuous	1.1289	1.6098
FallTwig	continuous	1.5285	2.0451
CoveMoss	continuous	2.1145	1.4787
CoveHerb	continuous	3.2550	1.2529
ReflLux	continuous	2.3618	1.5367

Table B.4: Summary of the 6 environmental variables of the Dutch dune spider data set. The variables are: *WaterCon,* the percentage of dry mass. *BareSand*, the percentage over of bare sand. *FallTwig*, the percentage over of fallen leaves and twigs. *CoveMoss*, the percentage cover of the moss layer. *CoveHerb*, the percentage cover of the herb layer. *RefLux*, the reflection of the soil surface with cloudless sky.

B.3 The orbited mite data set

	Min	Q1	Median	Mean	Q3	Max
Brachy	0.0000	3.0000	5.0000	8.7970	12.0000	42.0000
PHTH	0.0000	0.0000	0.0000	1.2900	2.0000	8.0000
HPAV	0.0000	4.0000	7.0000	8.5940	12.0000	37.0000
RARD	0.0000	0.0000	0.0000	1.2320	1.0000	13.0000
SSTR	0.0000	0.0000	0.0000	0.3188	0.0000	6.0000
Protopl	0.0000	0.0000	0.0000	0.3768	0.0000	13.0000
MEGR	0.0000	0.0000	1.0000	2.2170	3.0000	17.0000
MPRO	0.0000	0.0000	0.0000	0.1594	0.0000	2.0000
TVIE	0.0000	0.0000	0.0000	0.8406	1.0000	7.0000
HMIN	0.0000	0.0000	0.0000	4.9860	5.0000	36.0000
HMIN2	0.0000	0.0000	0.0000	1.9860	3.0000	20.0000
NPRA	0.0000	0.0000	1.0000	1.9130	3.0000	10.0000
TVEL	0.0000	0.0000	3.0000	9.1880	19.0000	42.0000
ONOV	0.0000	5.0000	11.0000	17.5200	25.0000	73.0000
SUCT	0.0000	8.0000	14.0000	17.2000	24.0000	63.0000
LCIL	0.0000	1.0000	12.0000	25.2900	44.0000	138.0000
Oribatl1	0.0000	0.0000	0.0000	1.9130	3.0000	17.0000
Ceratoz1	0.0000	0.0000	1.0000	1.3040	2.0000	5.0000
PWIL	0.0000	0.0000	0.0000	1.1010	1.0000	8.0000
Galumna1	0.0000	0.0000	0.0000	0.9710	1.0000	8.0000
Stgncrs2	0.0000	0.0000	0.0000	0.7391	0.0000	9.0000
HRUF	0.0000	0.0000	0.0000	0.2319	0.0000	3.0000
Trhypch1	0.0000	0.0000	0.0000	2.5510	2.0000	29.0000
PPEL	0.0000	0.0000	0.0000	0.1739	0.0000	3.0000
NCOR	0.0000	0.0000	1.0000	1.1450	2.0000	7.0000
SLAT	0.0000	0.0000	0.0000	0.4058	0.0000	8.0000
FSET	0.0000	0.0000	0.0000	1.8840	2.0000	12.0000
Lepidzts	0.0000	0.0000	0.0000	0.1739	0.0000	3.0000
Eupelops	0.0000	0.0000	0.0000	0.6522	1.0000	4.0000
Miniglmn	0.0000	0.0000	0.0000	0.2464	0.0000	5.0000
LRUG	0.0000	0.0000	4.0000	10.4200	18.0000	57.0000
PLAG2	0.0000	0.0000	0.0000	0.8116	1.0000	9.0000
Ceratoz3	0.0000	0.0000	0.0000	1.3190	2.0000	9.0000
Oppiminu	0.0000	0.0000	0.0000	1.1300	2.0000	9.0000
Trimalc2	0.0000	0.0000	0.0000	1.6230	0.0000	25.0000

Table B.5: Summary of the abundance of 35 orbited mite species. Min represents the minimum, Q1 represents the first quantile, Median is the second quantile, Mean is the average, Q3 represents the third quantile and Max represents the maximum.

	Type	Description
SubsDens	continuous	Mean=39.0971; Standard deviation=11.9287
WatrCont	continuous	Mean=404.602; Standard deviation=134.0881
Substrate	category	levels: Sphagn1, Sphagn2, Sphagn3, Sphagn4, Litter, Barepeat, Interface.
Shrub	category	levels: None, Few, Many.
Topo	category	levels: Blanket, Hummotck.

Table B.6: Overview of the substratum type. *SubsDens*: the density of the substratum, *WatrCont*: the water content of the substrate, *Substrate*: the substratum type which has 7 classes, *Shrub*: the coverage density of the shrub and *Topo*: Microtopograhy, a factor with 2 levels.

B.4 Human infant gut microbiome data set

	BF	Infant_Formula	Oat	Barley	Rye	Buckwheat_Millet	Cereal	Root_Veg	Veg	Eggs	Soy_Prod	Milk_Prod	Meat	Fish	Solid_Food	Age_at_Collection	Subject_ID
G36449	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	62	E001463
G36034	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	82	E001463
G36993	1	1	0	0	0	0	1	1	1	0	0	0	0	0	1	124	E001463
G35523	1	1	1	0	0	0	1	1	1	0	0	0	1	0	1	153	E001463
G36450	1	1	1	0	1	0	1	1	1	0	0	0	1	1	1	187	E001463
G36028	1	1	1	0	1	0	1	1	1	0	0	0	1	1	1	213	E001463
G36029	1	1	1	1	1	0	1	1	1	0	0	0	1	1	1	243	E001463
G35524	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	276	E001463
G36451	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	303	E001463
G35424	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	366	E001463
G35522	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	396	E001463
G36452	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	427	E001463
G36025	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	457	E001463
G36030	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	487	E001463
G35521	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	518	E001463
G36453	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	537	E001463
G36026	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	577	E001463
G35423	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	608	E001463
G35867	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	638	E001463
G36037	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	668	E001463
G35531	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	701	E001463
G35532	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	738	E001463
G35534	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	796	E001463
G36847	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	853	E001463
G36839	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	883	E001463
G36837	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	918	E001463
G36829	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	943	E001463
G36834	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	972	E001463
G36835	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1006	E001463
G36811	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1028	E001463
G35881	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1062	E001463
G35533	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1136	E001463

Table B.7: Th dietary intake of infant ‘E001463’. 0 indicates the infant does not take the diet and 1 indicates the infant takes the diet.

Appendix C

Graph

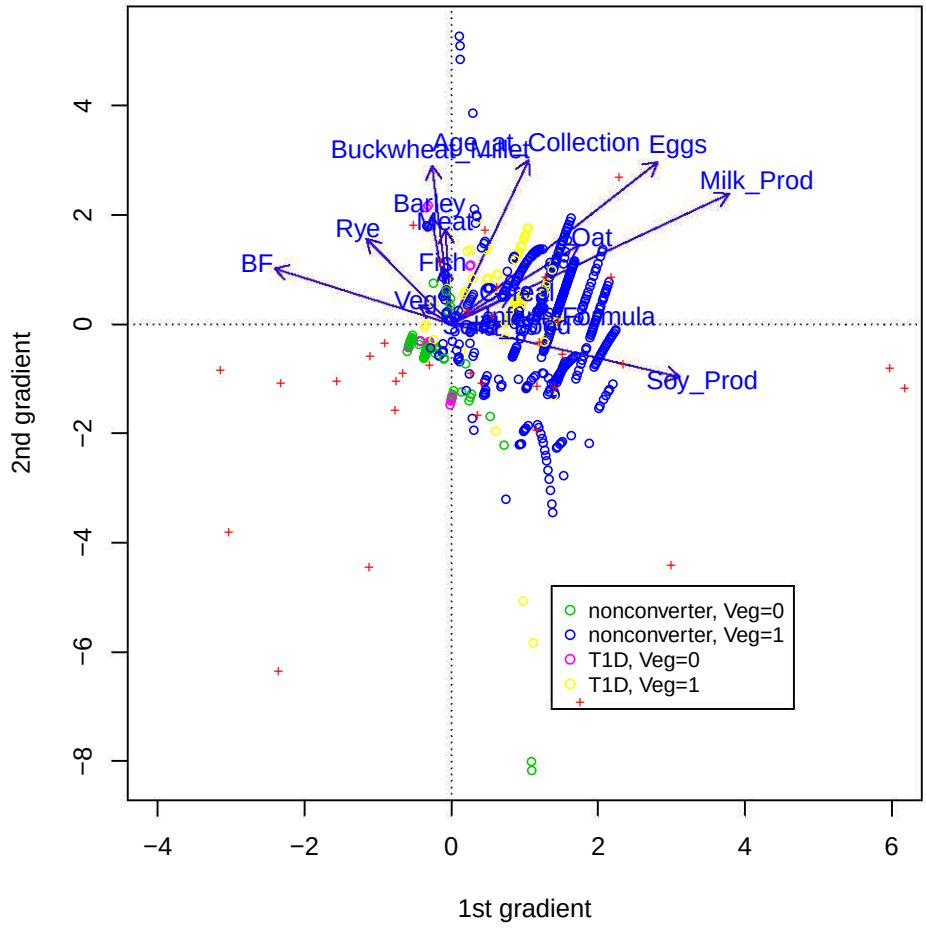


Figure C.1: Ordination diagram of ZINB based approach. Arrows represent the diet intake of the infant's, crossed signs represents microbiome families and circles are for the samples. The samples are coloured according to whether vegetable is in the diet.

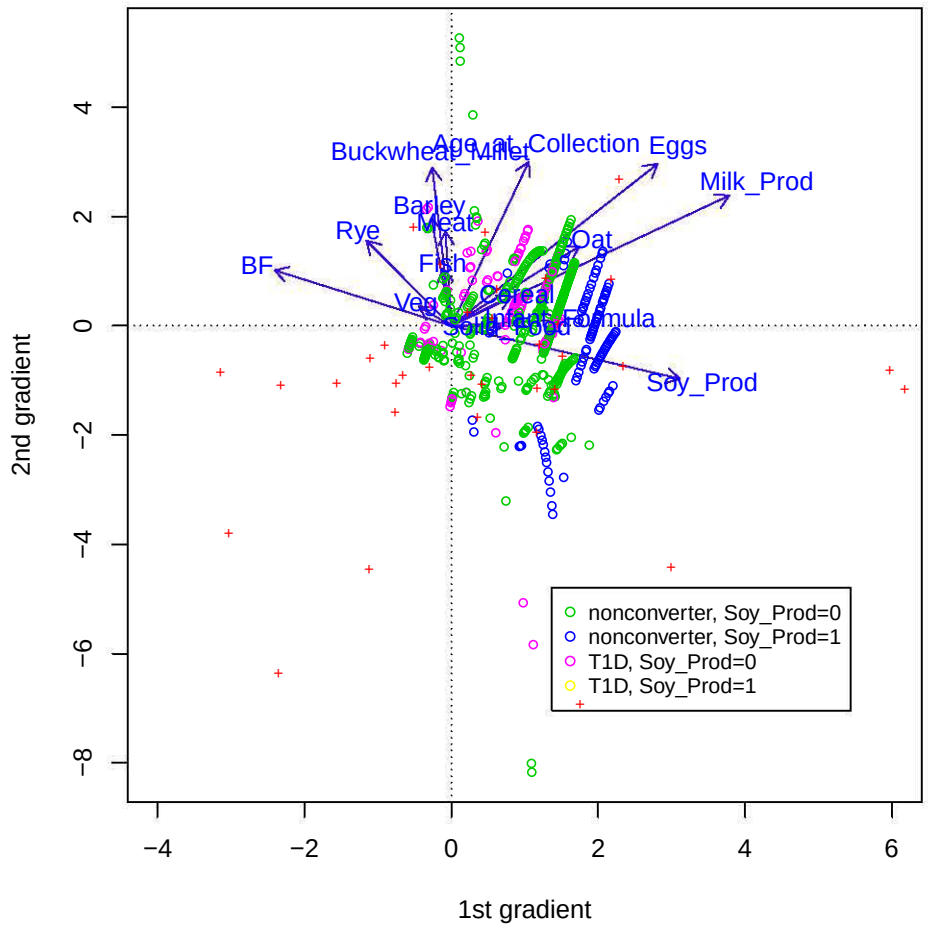


Figure C.2: Ordination diagram of ZINB based approach. Arrows represent the diet intake of the infant's, crossed signs represents microbiome families and circles are for the samples. The samples are coloured according to whether soy produce is in the diet.

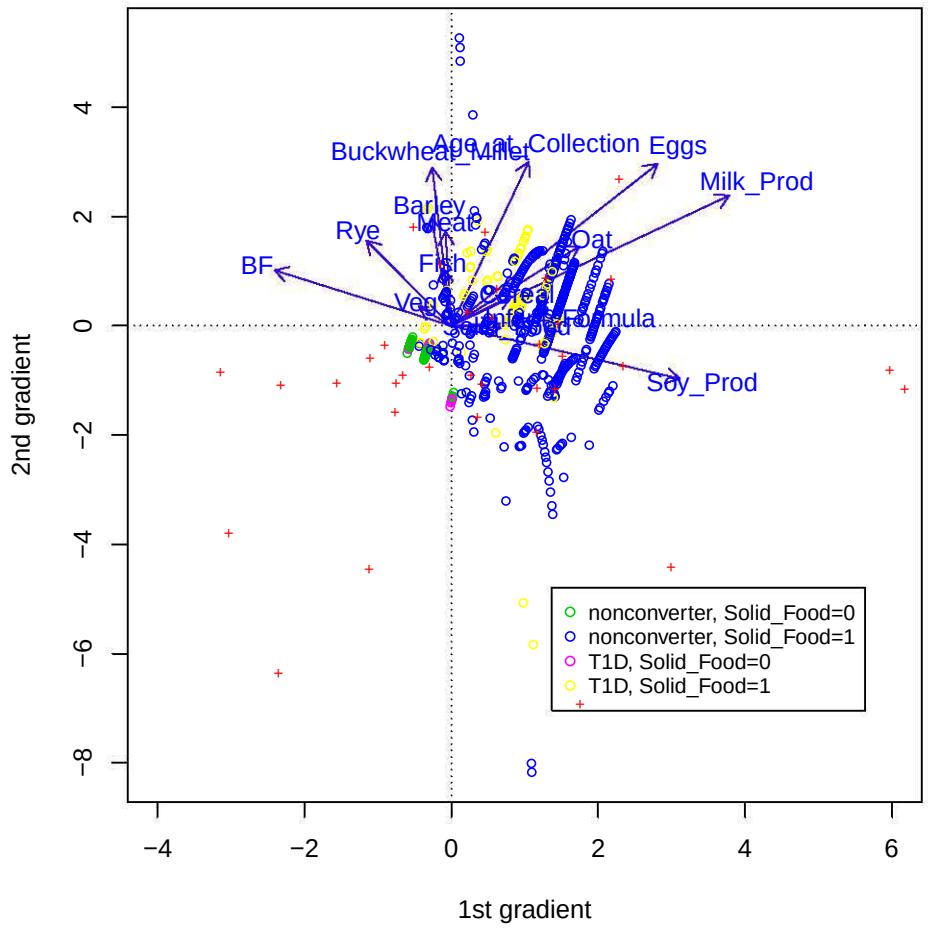


Figure C.3: Ordination diagram of ZINB based approach. Arrows represent the diet intake of the infant's, crossed signs represents microbiome families and circles are for the samples. The samples are coloured according to whether the infant receive solid food.

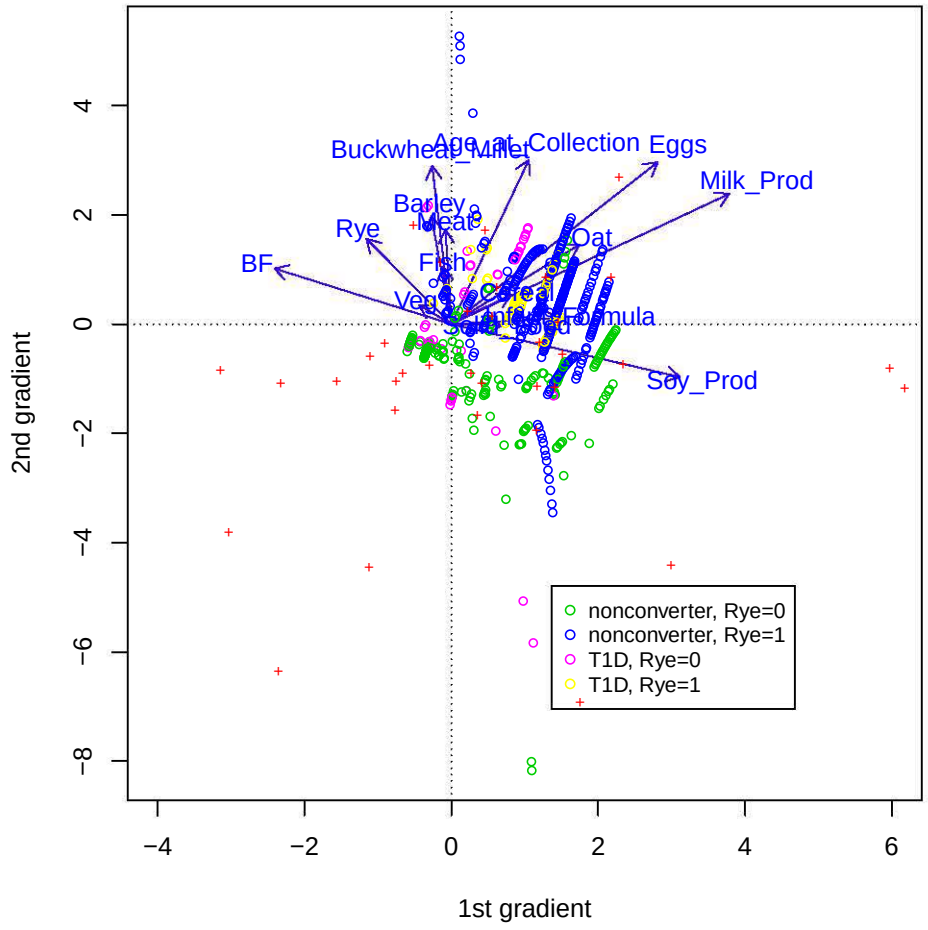


Figure C.4: Ordination diagram of ZINB based approach. Arrows represent the diet intake of the infant's, crossed signs represents microbiome families and circles are for the samples. The samples are coloured according to whether rye is in the diet.

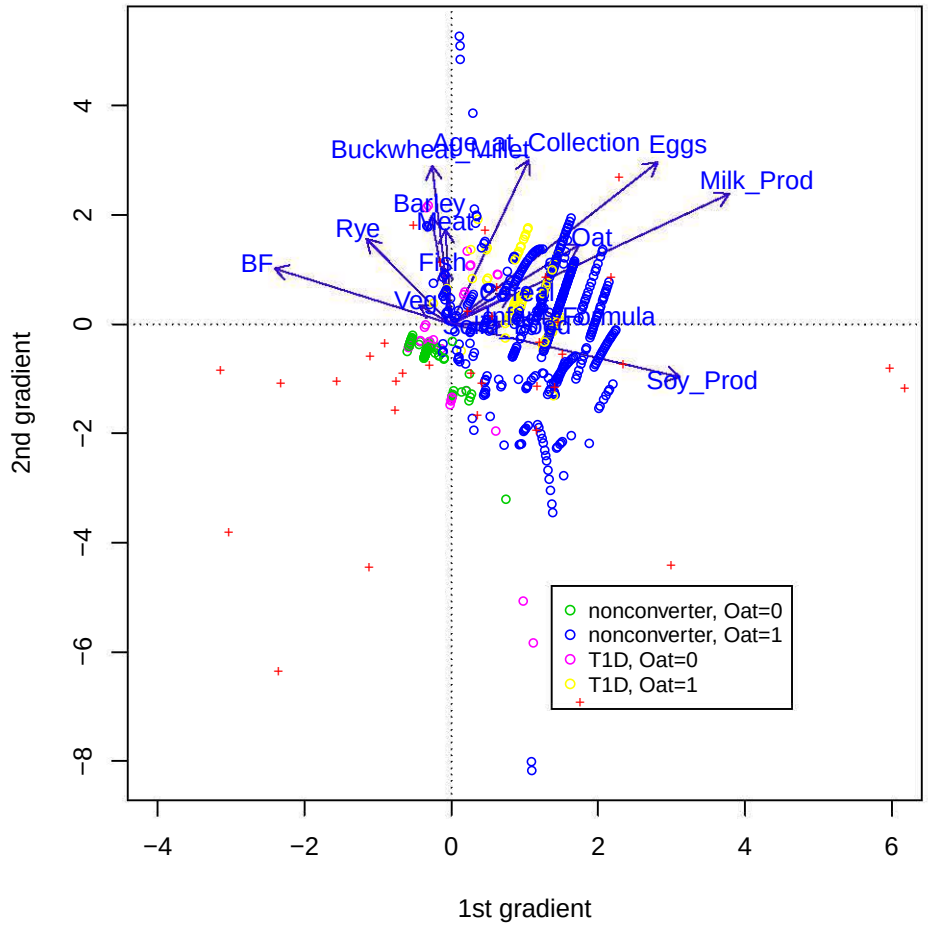


Figure C.5: Ordination diagram of ZINB based approach. Arrows represent the diet intake of the infant's, crossed signs represents microbiome families and circles are for the samples. The samples are coloured according to whether oat is in the diet.

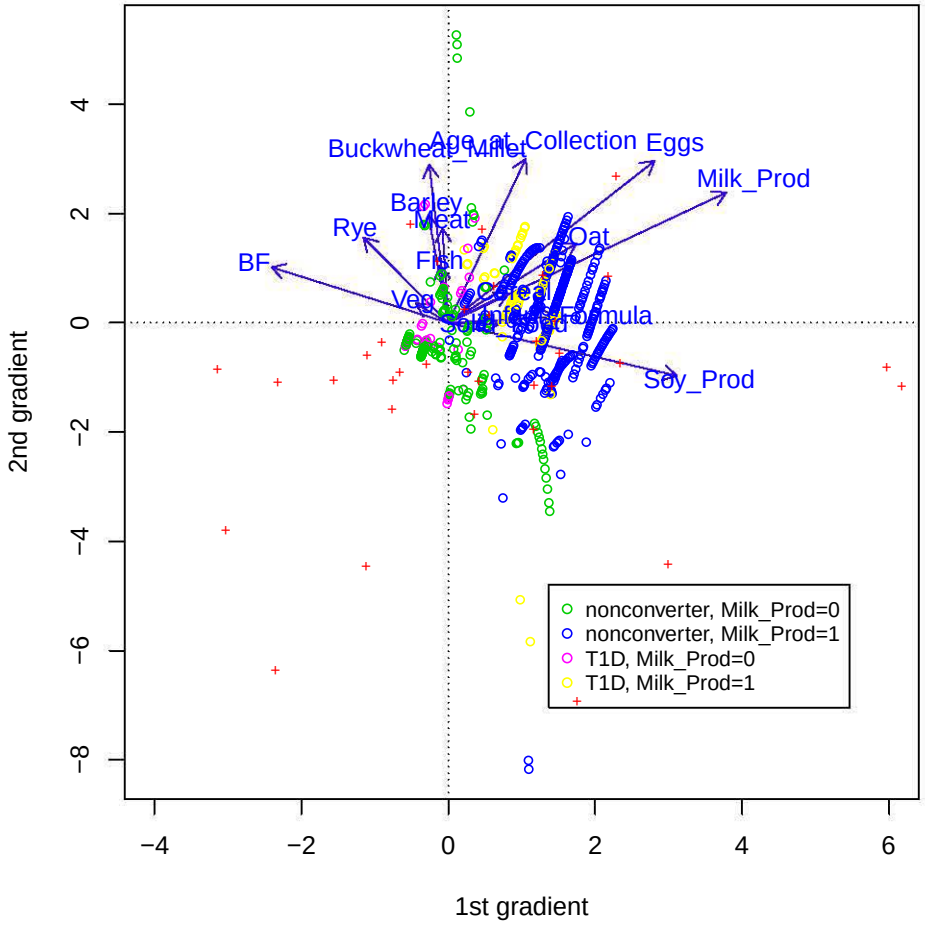


Figure C.6: Ordination diagram of ZINB based approach. Arrows represent the diet intake of the infant's, crossed signs represents microbiome families and circles are for the samples. The samples are coloured according to whether milk product is in the diet.

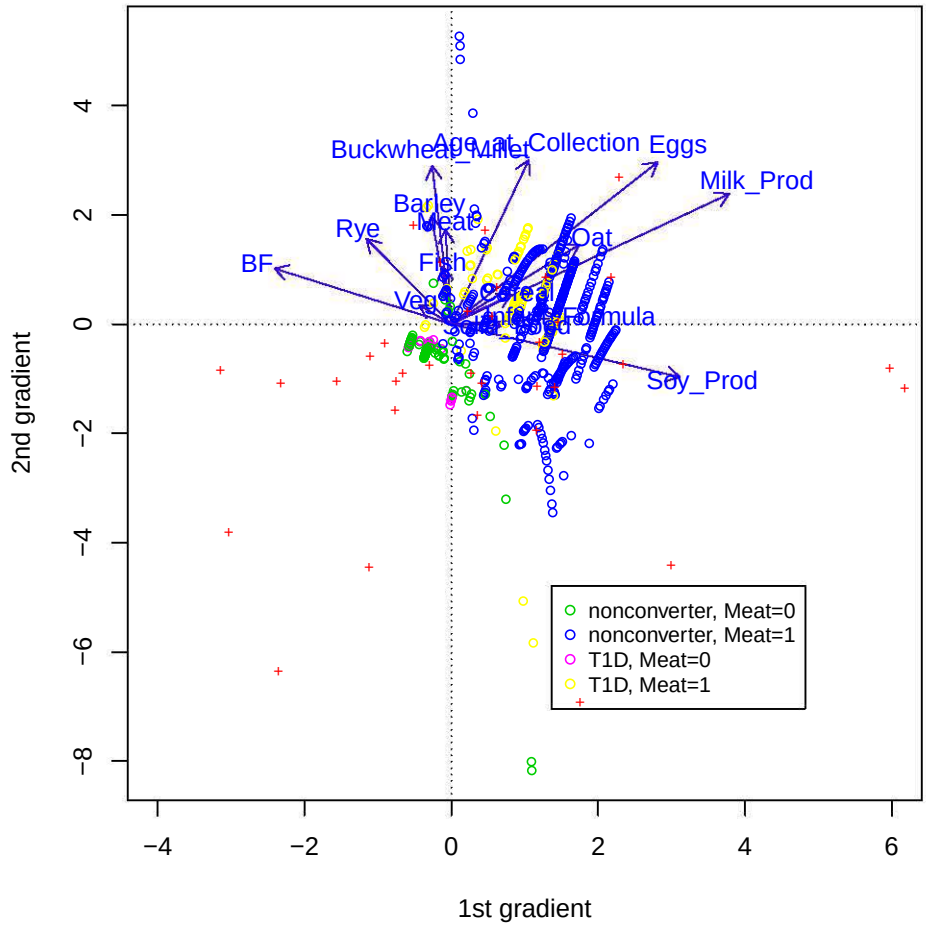


Figure C.7: Ordination diagram of ZINB based approach. Arrows represent the diet intake of the infant's, crossed signs represents microbiome families and circles are for the samples. The samples are coloured according to whether meat is in the diet.

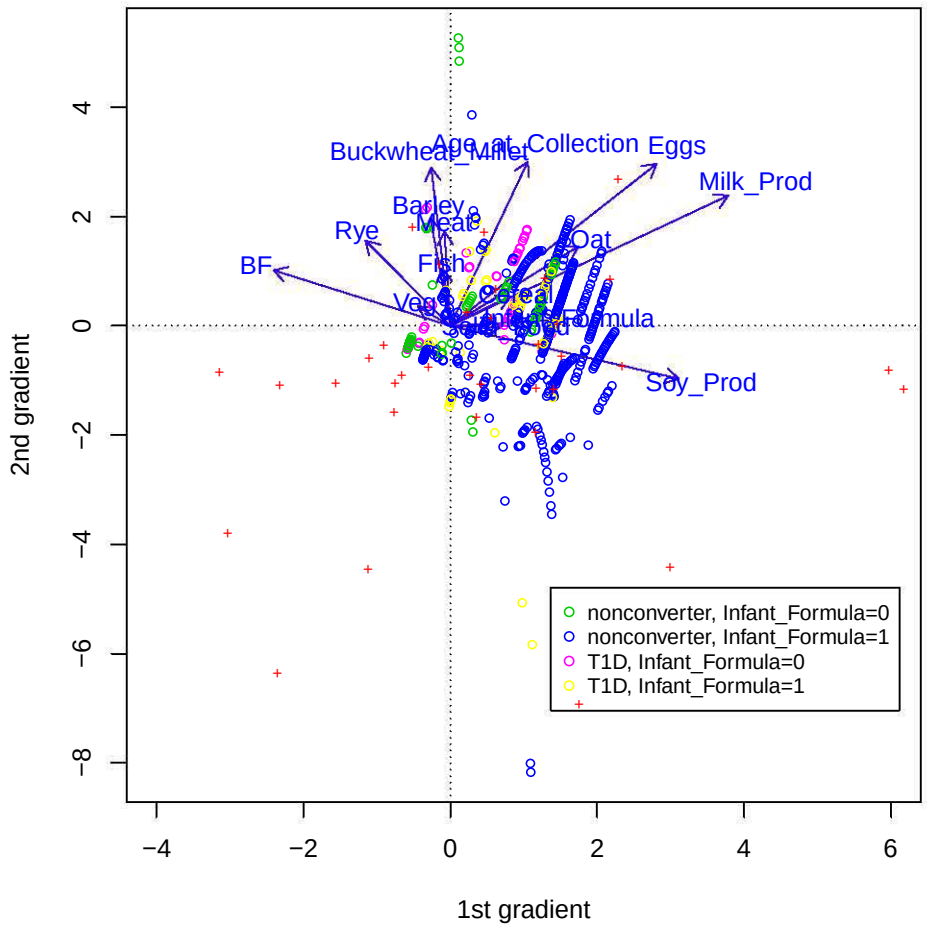


Figure C.8: Ordination diagram of ZINB based approach. Arrows represent the diet intake of the infant's, crossed signs represents microbiome families and circles are for the samples. The samples are coloured according to whether infant formula is in the diet.

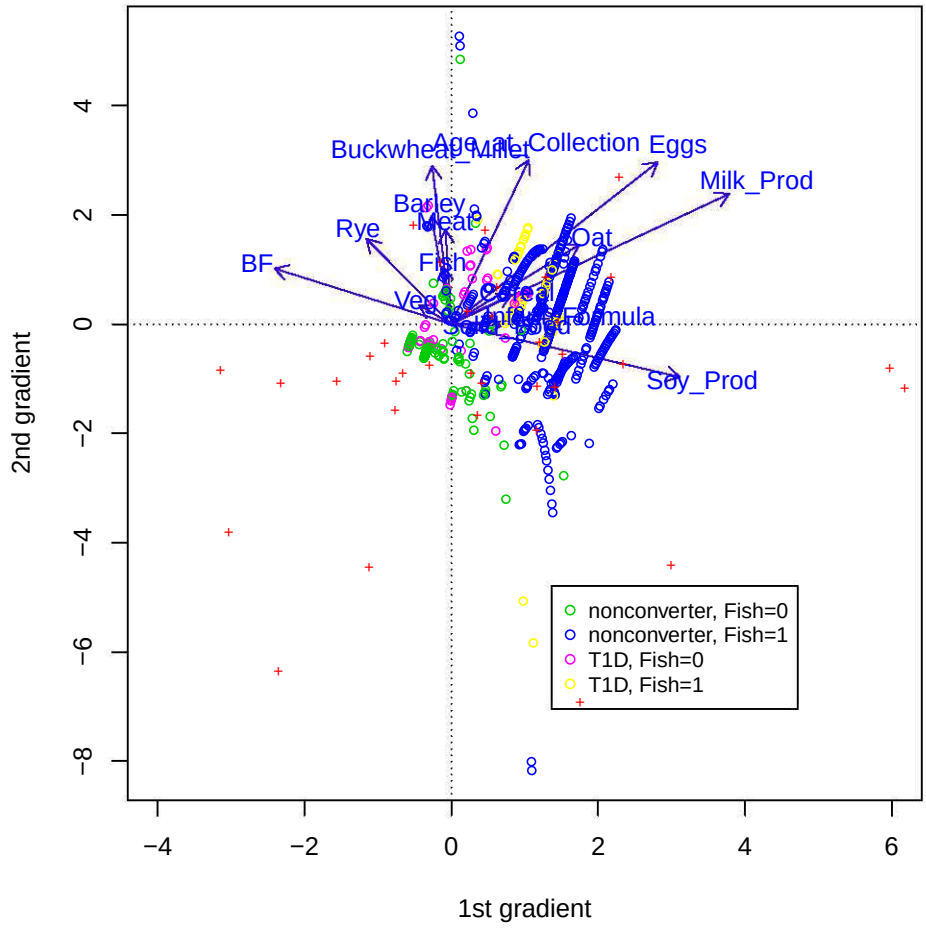


Figure C.9: Ordination diagram of ZINB based approach. Arrows represent the diet intake of the infant's, crossed signs represents microbiome families and circles are for the samples. The samples are coloured according to whether fish is in the diet.

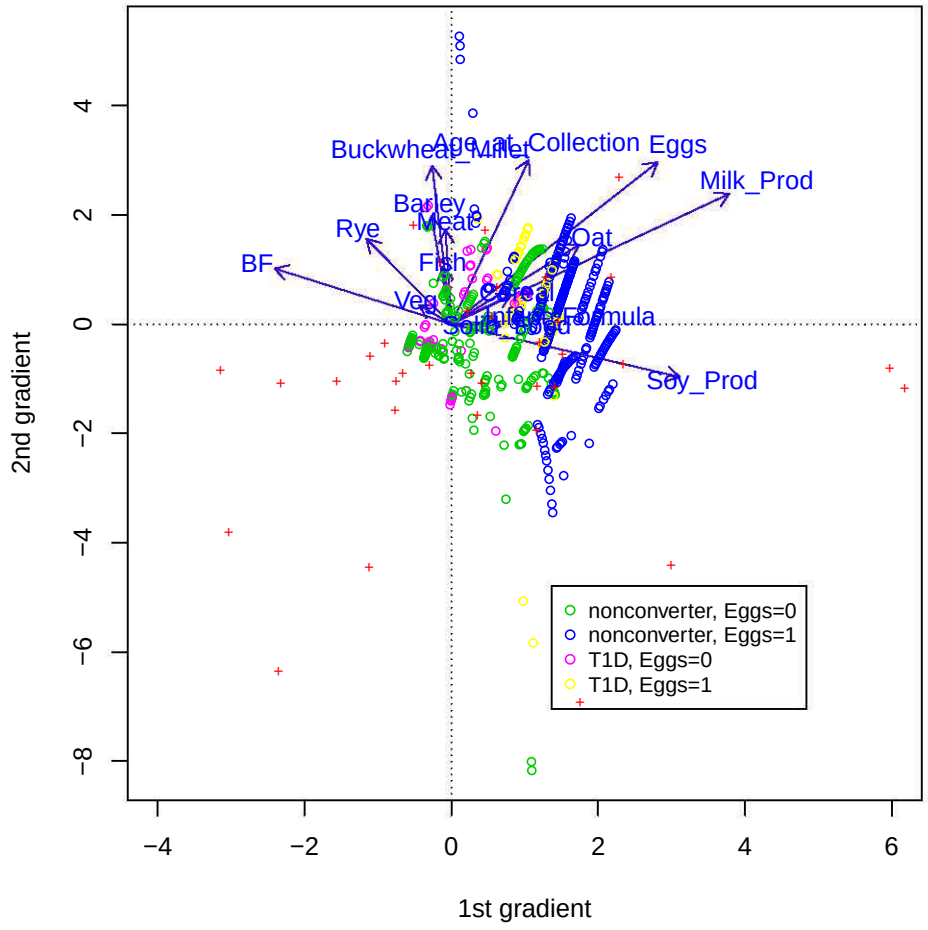


Figure C.10: Ordination diagram of ZINB based approach. Arrows represent the diet intake of the infant's, crossed signs represents microbiome families and circles are for the samples. The samples are coloured according to whether eggs is in the diet.

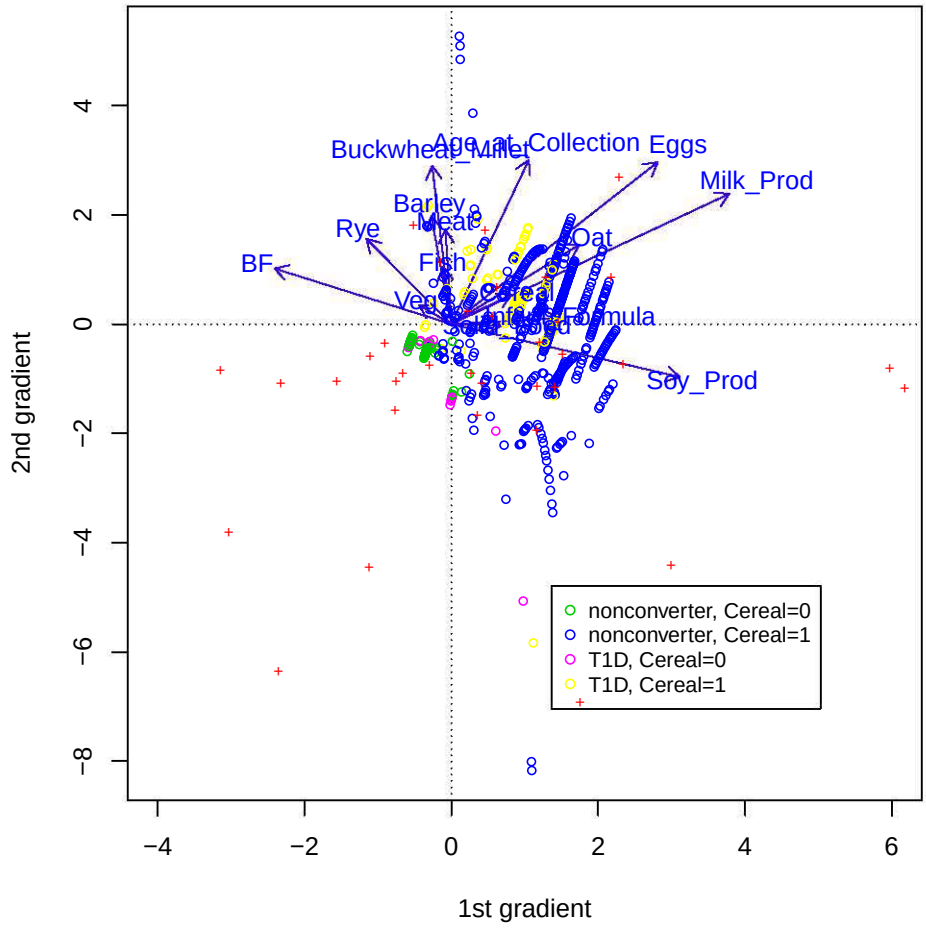


Figure C.11: Ordination diagram of ZINB based approach. Arrows represent the diet intake of the infant's, crossed signs represents microbiome families and circles are for the samples. The samples are coloured according to whether cereal is in the diet.

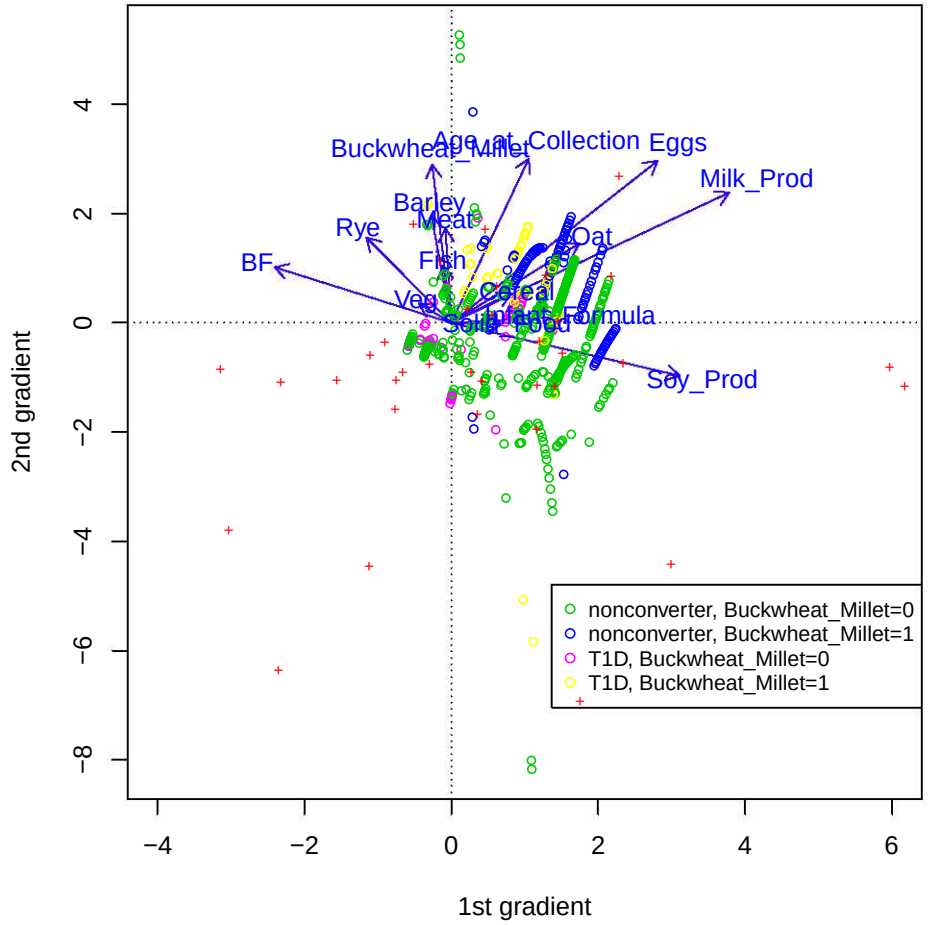


Figure C.12: Ordination diagram of ZINB based approach. Arrows represent the diet intake of the infant's, crossed signs represents microbiome families and circles are for the samples. The samples are coloured according to whether buckwheat millet is in the diet.

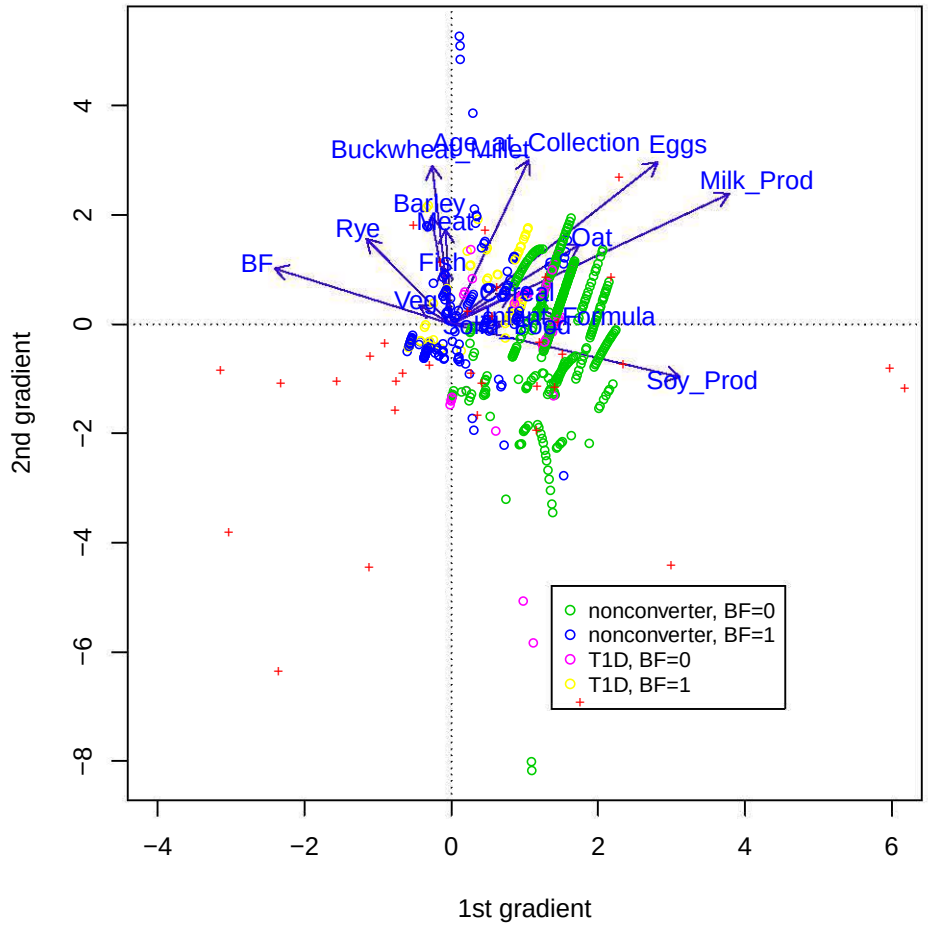


Figure C.13: Ordination diagram of ZINB based approach. Arrows represent the diet intake of the infant's, crossed signs represents microbiome families and circles are for the samples. The samples are coloured according to whether the infant is breastfed.

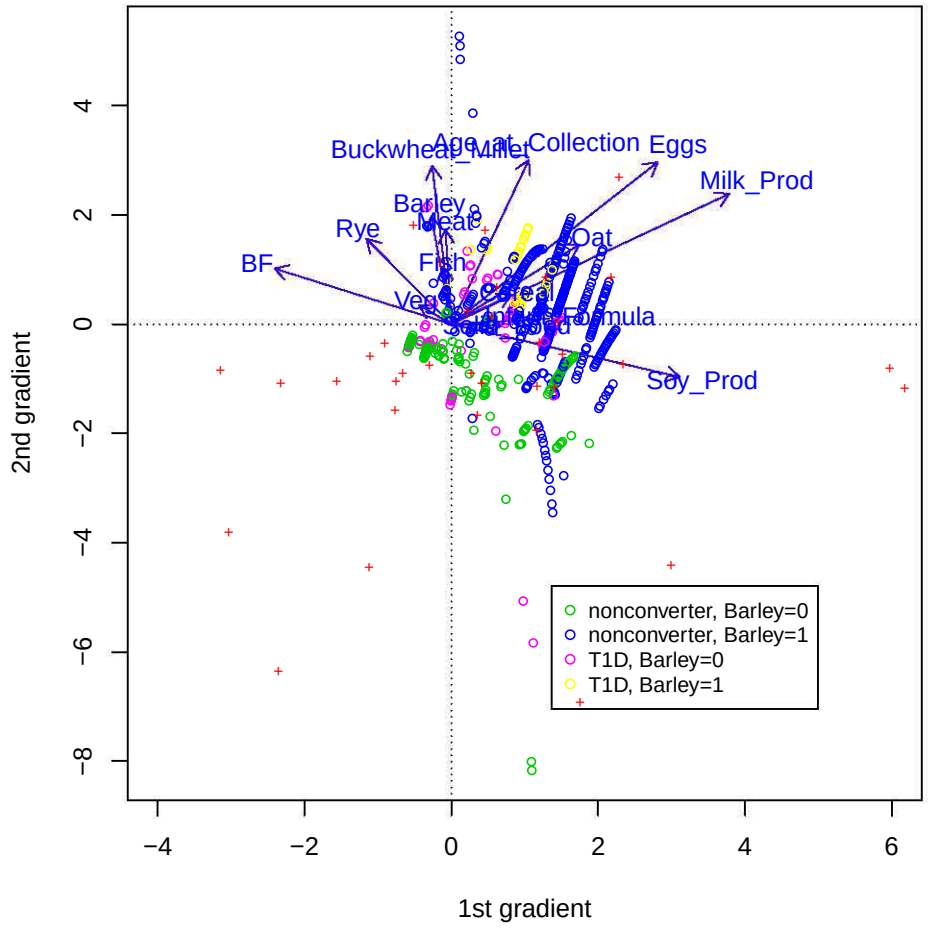


Figure C.14: Ordination diagram of ZINB based approach. Arrows represent the diet intake of the infant's, crossed signs represents microbiome families and circles are for the samples. The samples are coloured according to whether barley is in the diet.